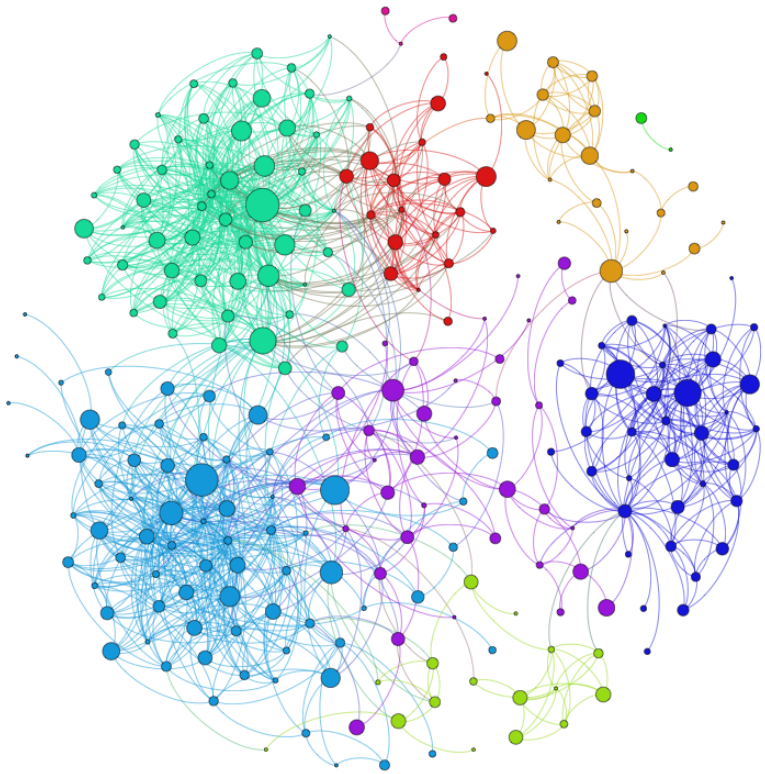
Data Analysis for Economics

ä Modern Introduction

Jiaming Mao

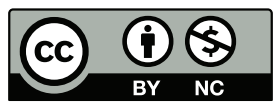
Xiamen University



This version: Spring 2022

Contact: jmao@xmu.edu.cn

Course homepage: jiamingmao.github.io/data-analysis



Data are everywhere

	Cheese				
0.5/0.51 lb	**Cabot Vermont Cheddar**	0.51 lb	$7.99/lb	**$4.07**	
	Dairy				
1/1	**Friendship Lowfat Cottage Cheese** (16oz)		$2.89/ea	**$2.89**	
1/1	**Nature's Yoke Grade A Jumbo Brown Eggs** (1 dozen)		$1.49/ea	**$1.49**	
1/1	**Santa Barbara Hot Salsa, Fresh** (16oz)		$2.69/ea	**$2.69**	
1/1	**Stonyfield Farm Organic Lowfat Plain Yogurt** (32oz)		$3.59/ea	**$3.59**	
	Fruit				
3/3	**Anjou Pears** (Farm Fresh, Med)	1.76 lb	$2.49/lb	**$4.38**	
2/2	**Cantaloupe** (Farm Fresh, Med)		$2.00/ea	**$4.00**	**S**
	Grocery				
1/1	**Fantastic World Foods Organic Whole Wheat Couscous** (12oz)		$1.99/ea	**$1.99**	
1/1	**Garden of Eatin' Blue Corn Chips** (9oz)		$2.49/ea	**$2.49**	
1/1	**Goya Low Sodium Chickpeas** (15.5oz)		$0.89/ea	**$0.89**	
2/2	**Marcal 2-Ply Paper Towels, 90ct** (1ea)		$1.09/ea	**$2.18**	**T**
1/1	**Muir Glen Organic Tomato Paste** (6oz)		$0.99/ea	**$0.99**	
1/1	**Starkist Solid White Albacore Tuna in Spring Water** (6oz)		$1.89/ea	**$1.89**	

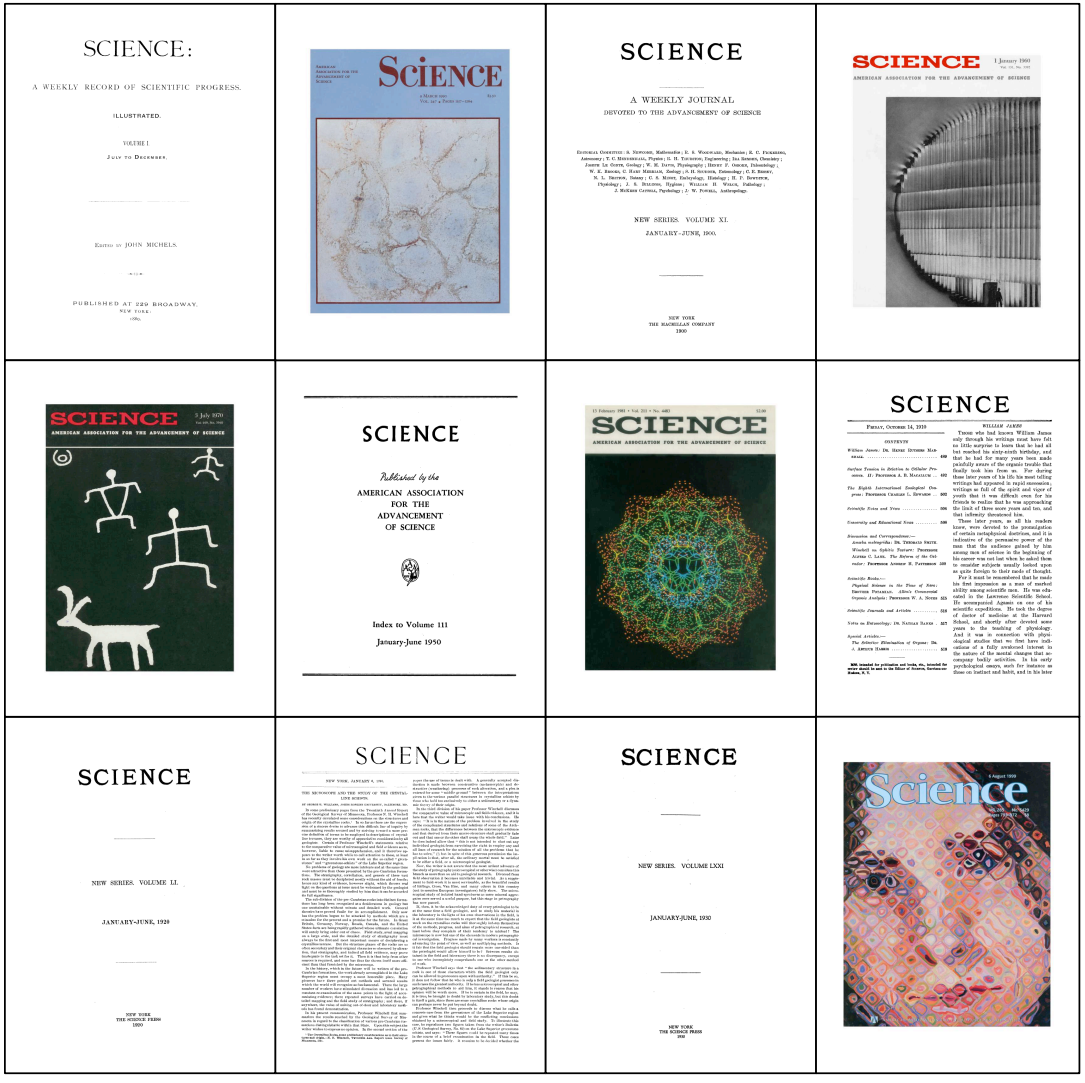
Purchase histories

Data are everywhere

Ikiru (1952)	UR	Foreign	
Junebug (2005)	R	Independent	
La Cage aux Folles (1979)	R	Comedy	
The Life Aquatic with Steve Zissou (2004)	R	Comedy	
Lock, Stock and Two Smoking Barrels (1998)	R	Action & Adventure	
Lost in Translation (2003)	R	Drama	
Love and Death (1975)	PG	Comedy	
The Manchurian Candidate (1962)	PG-13	Classics	
Memento (2000)	R	Thrillers	
Midnight Cowboy (1969)	R	Classics	

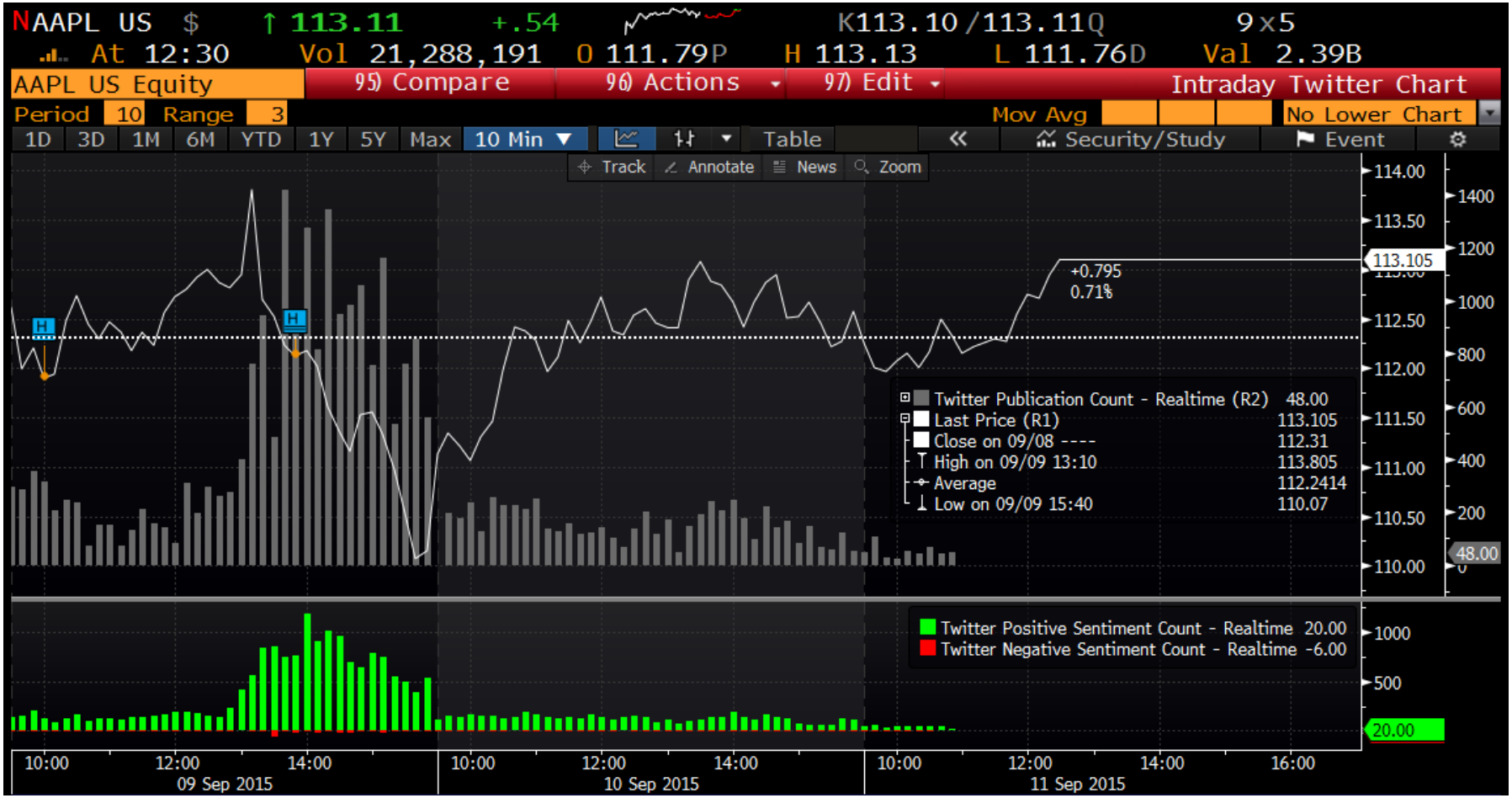
User ratings

Data are everywhere



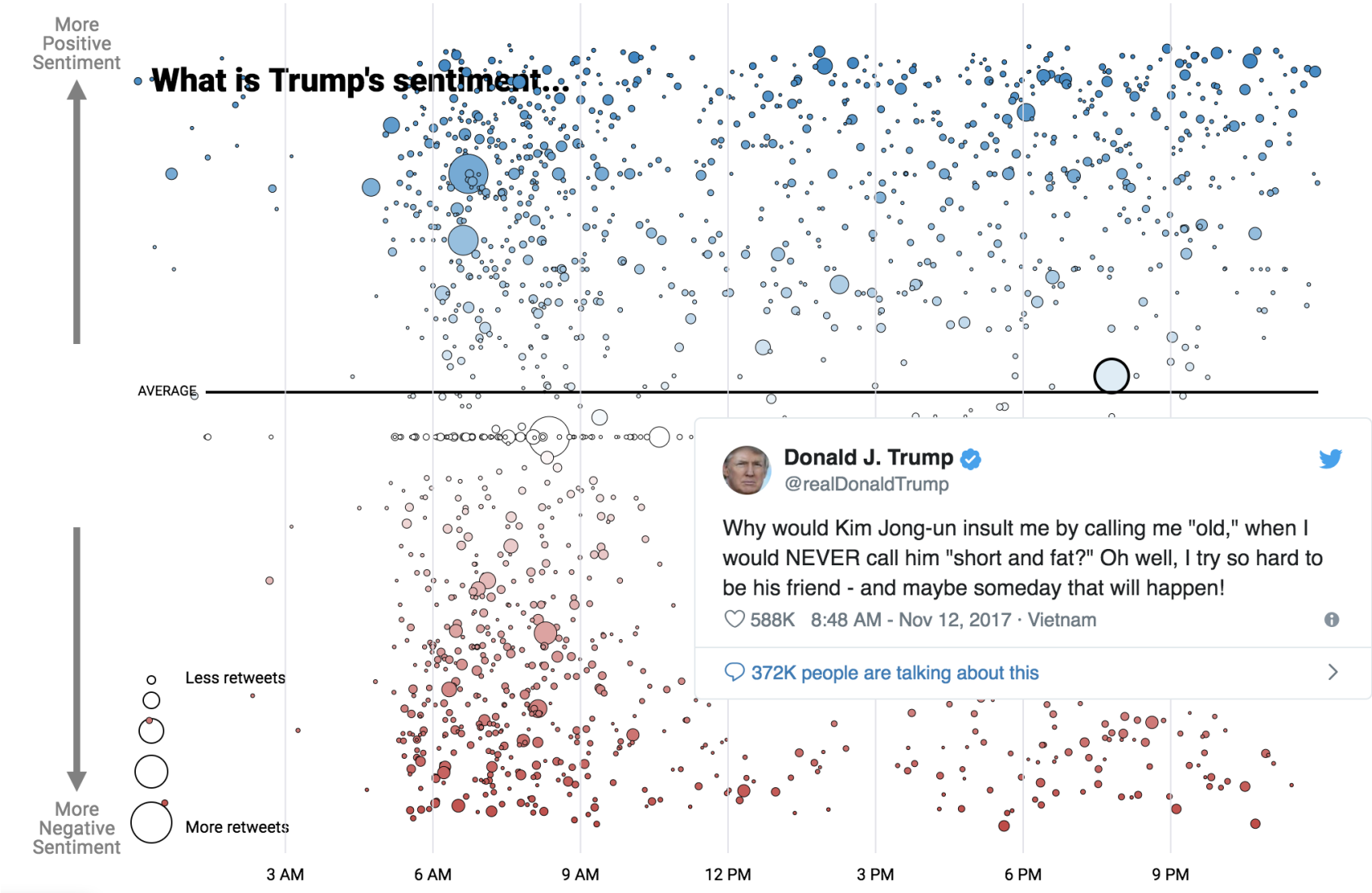
Document collections

Data are everywhere



Financial markets

Data are everywhere



Social networks

Data Science

"What's in a name? that which we call a rose,
By any other name would smell as sweet." – *Juliet*

Machine Learning $\rightarrow$ Statistics $\rightarrow$ Econometrics

- Along this spectrum, the focus moves from prediction and pattern discovery to inference about causality and the underlying mechanisms that generate the observed data.

Pattern Discovery

Classification









Which one is a chair?

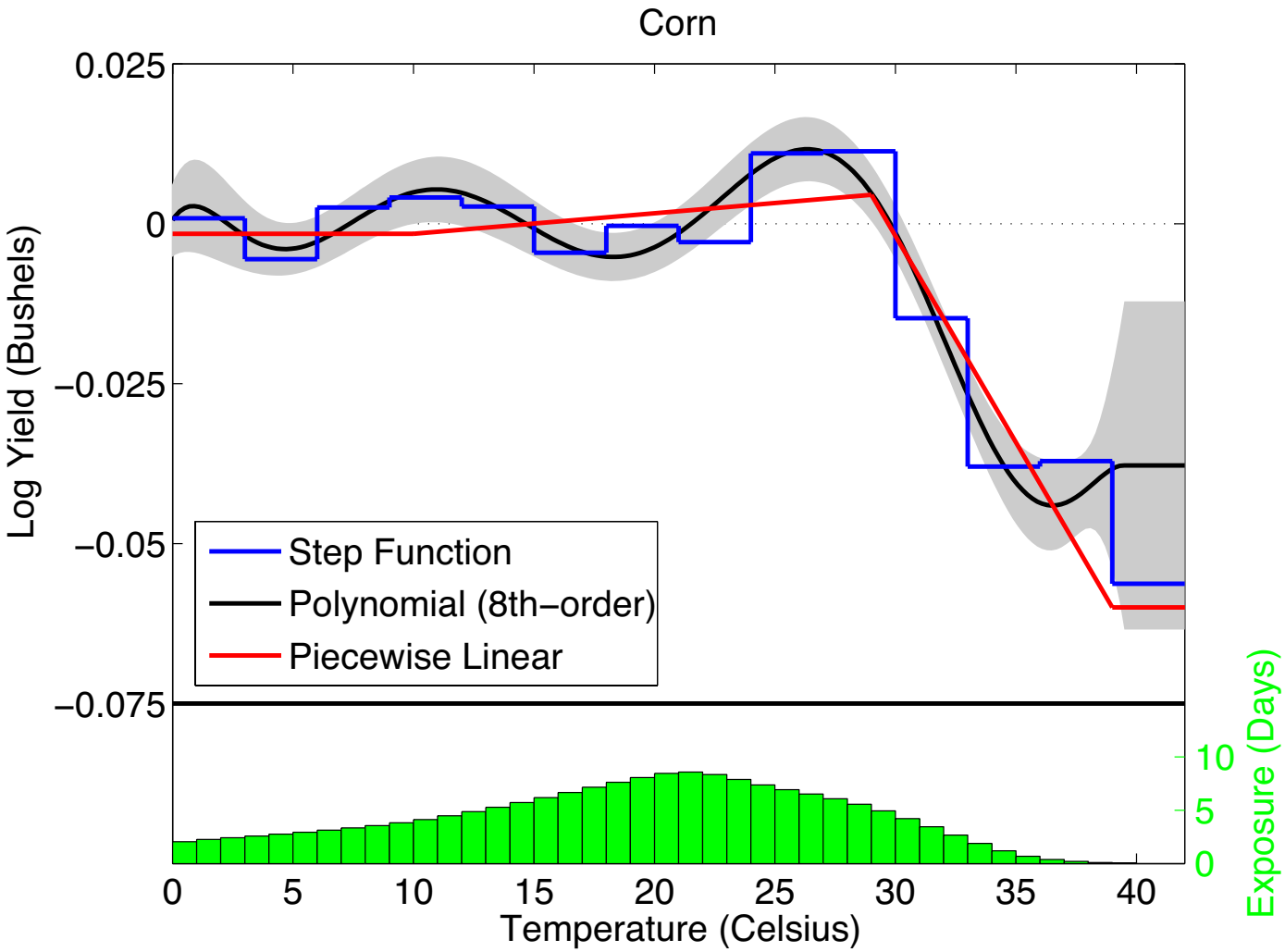
Pattern Discovery

Classification

- Which product will a consumer buy?
- Which market will a firm enter?
- Which political candidate will an individual vote for?

Pattern Discovery

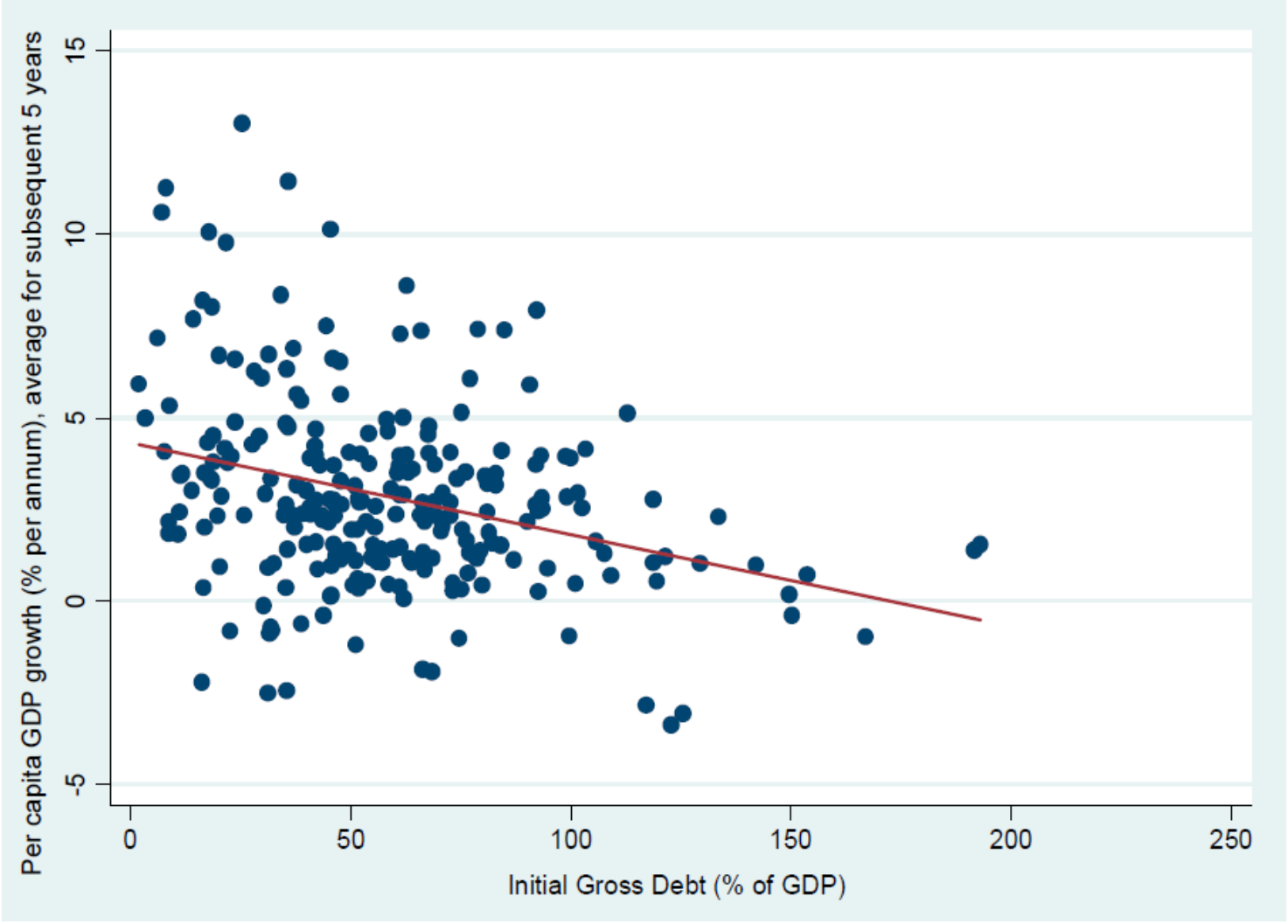
Regression



Temperature and crop yield

Pattern Discovery

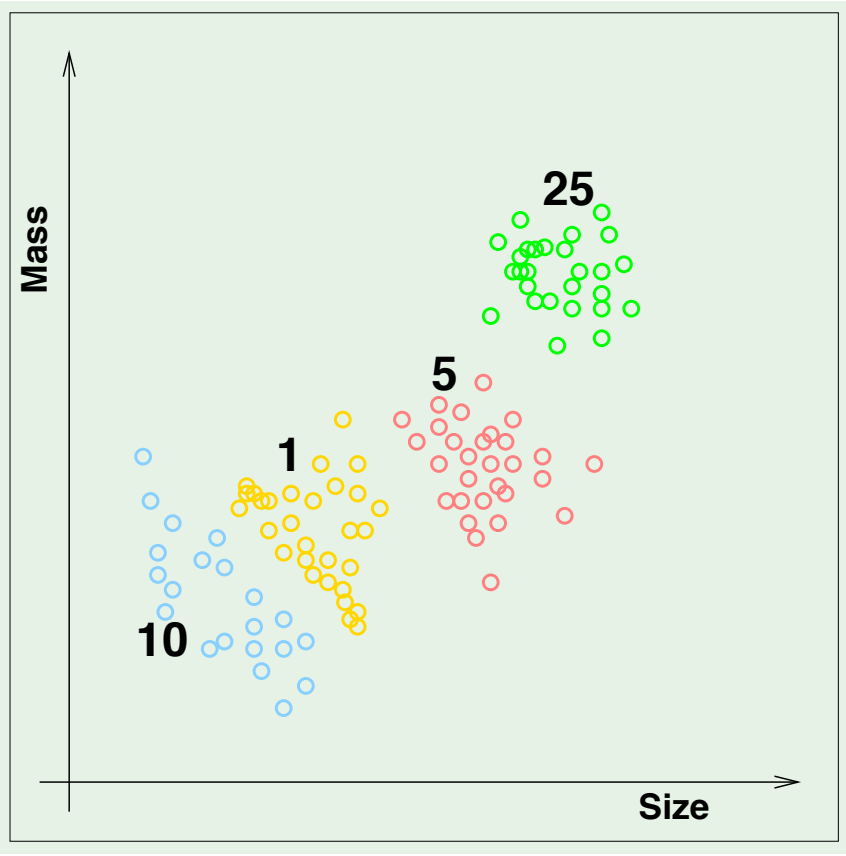
Regression

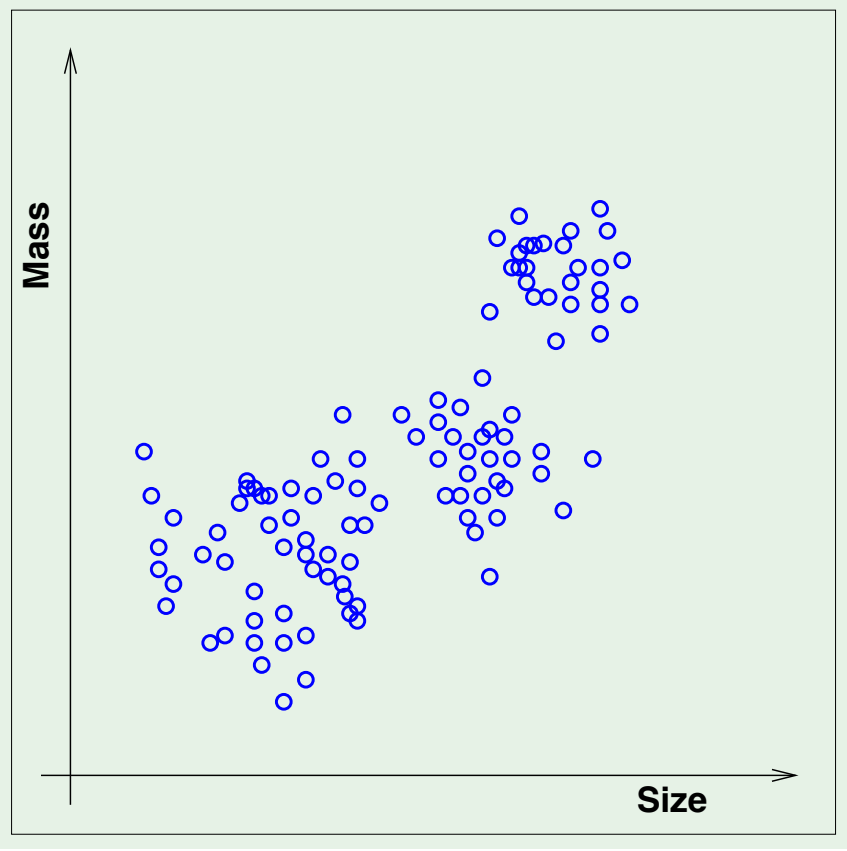


Government debt and GDP growth

Pattern Discovery

Unsupervised Learning



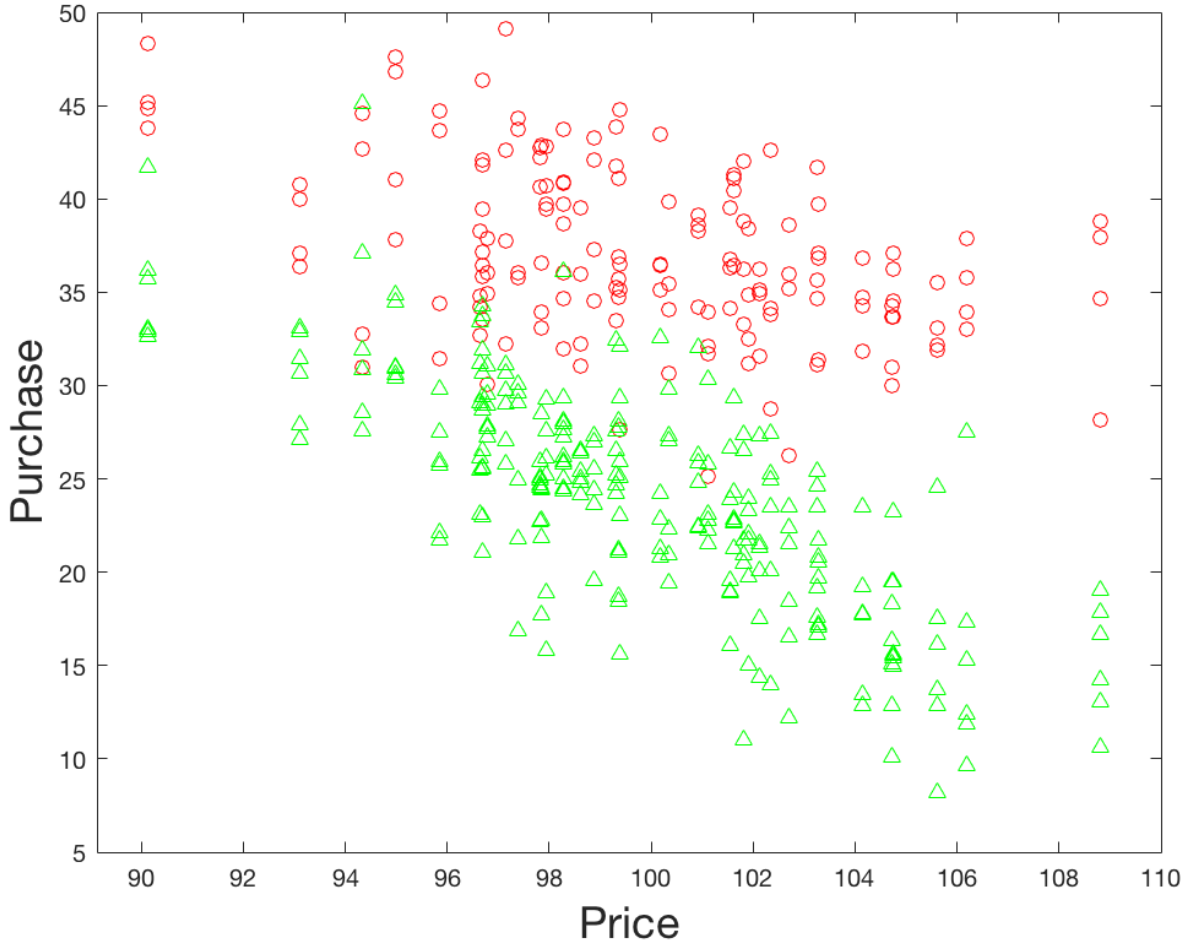


Vending machine coin recognition

Left: supervised learning; Right: unsupervised learning

Pattern Discovery

Unsupervised Learning

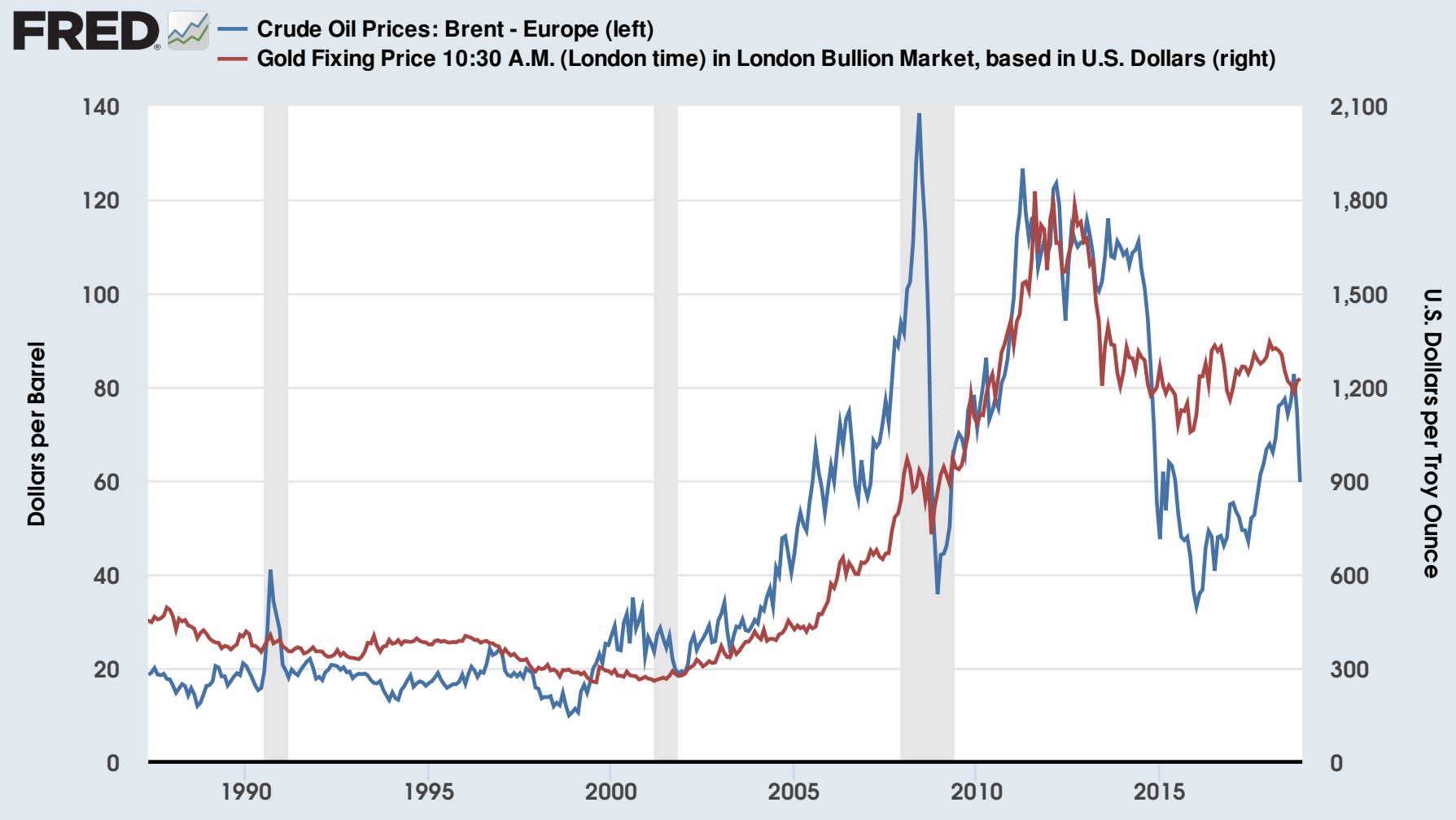


Consumer demand

Causal Inference

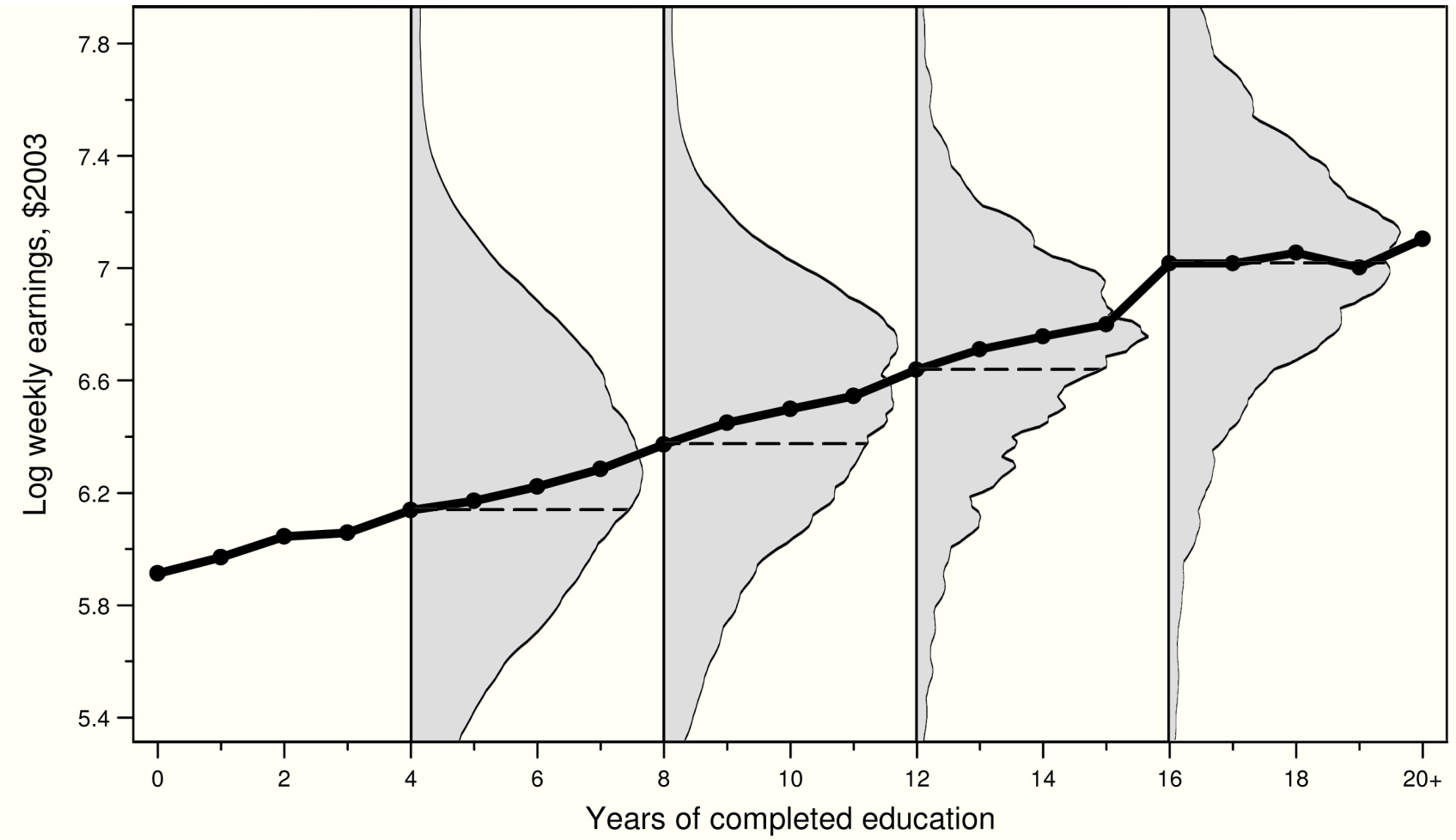
Learning patterns in the data is not enough – we want understanding.

Causal Inference



Do gold and oil prices cause each other to move or are their comovements caused by something else?

Causal Inference



Does receiving more education make you earn more?

Program Evaluation

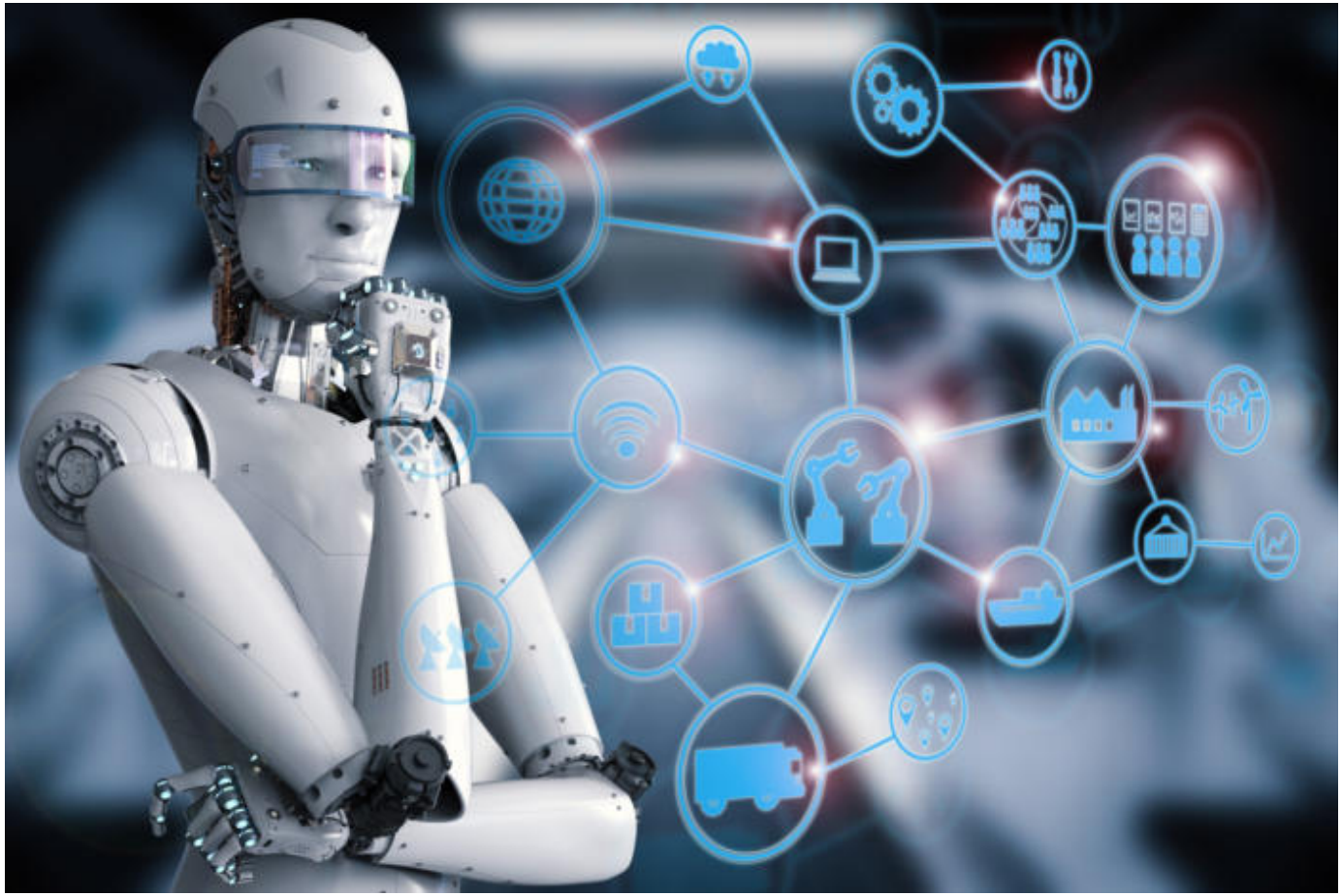
Evaluating and predicting the effects of government programs and economic policies is a central problem in applied economic research:

- Effect of worker training programs on employment
- Effect of income taxes on labor supply
- Effect of zoning regulations on housing prices
- Effect of environmental regulations on pollution emission
- ...

Artificial Intelligence

- Research on causal inference methodologies has taken on new importance with the development of artificial intelligence (AI).
- So far, progress in causal inference has been made mainly in developing methods to learn causal effects or estimate causal models from data based on our understanding of the underlying mechanisms.
- Models of causal mechanisms are developed by human experts.
  - Science progresses by formulating models of causal mechanisms, then conduct experiments or observational studies, and update the models based on their results.
- Building machines that can learn causal mechanisms without human experts would be the ultimate goal of artificial intelligence.

Artificial Intelligence

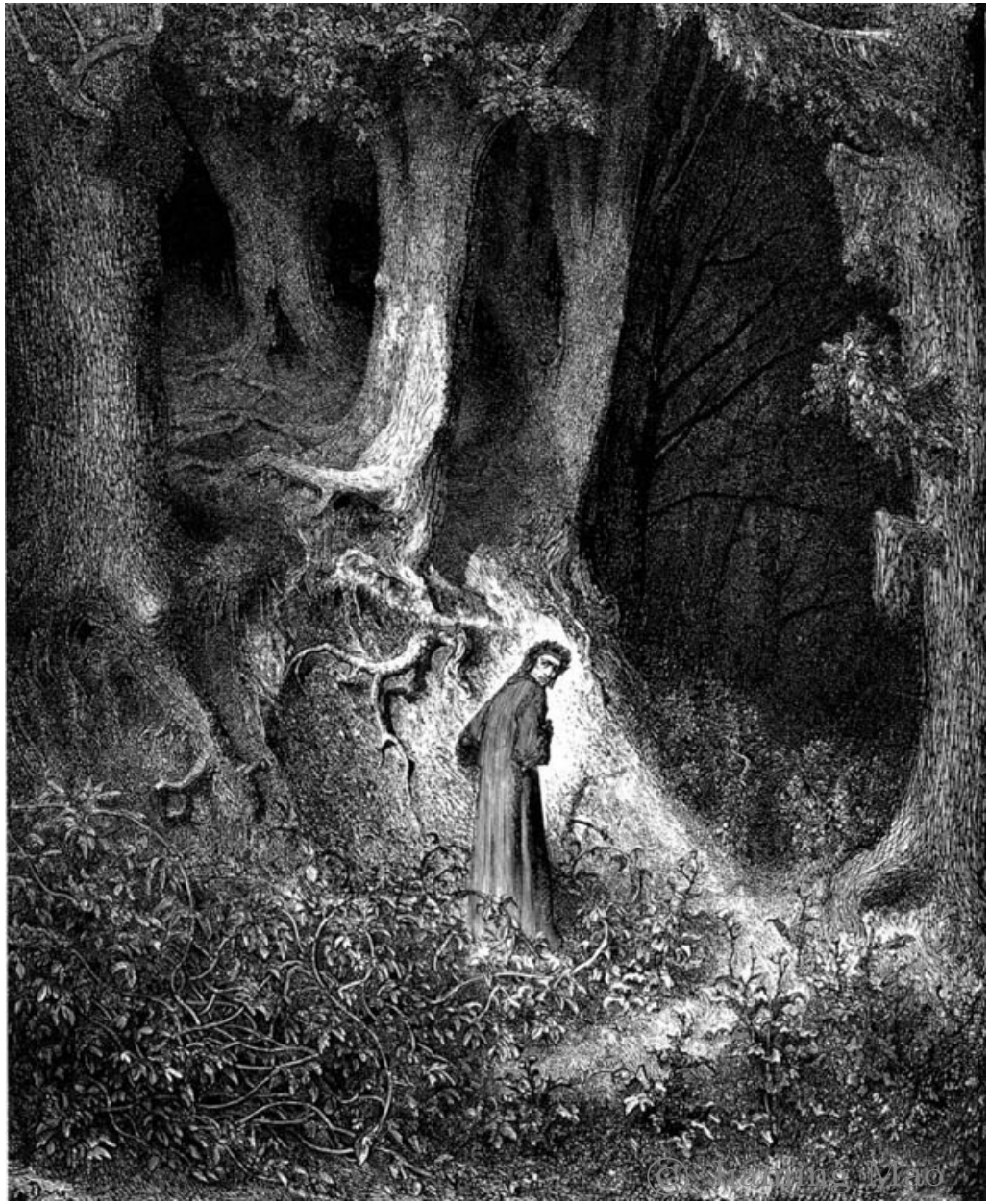


Road Map

1. Statistical Modeling
2. Causal Inference
3. Structural Estimation

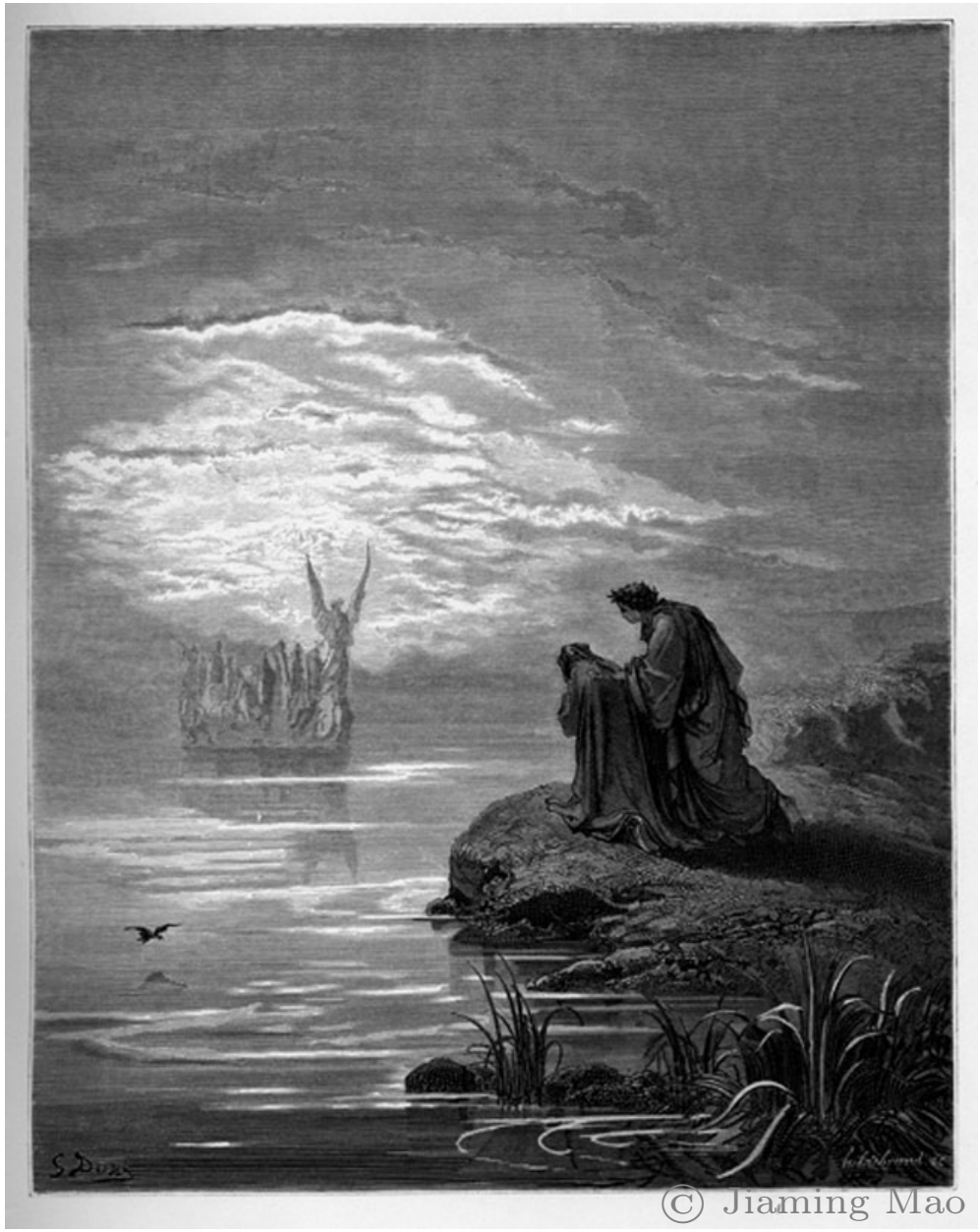
Road Map

Thematically, we follow the journey of a hero determined to seek knowledge from data, who departs the *forest of ignorance*,



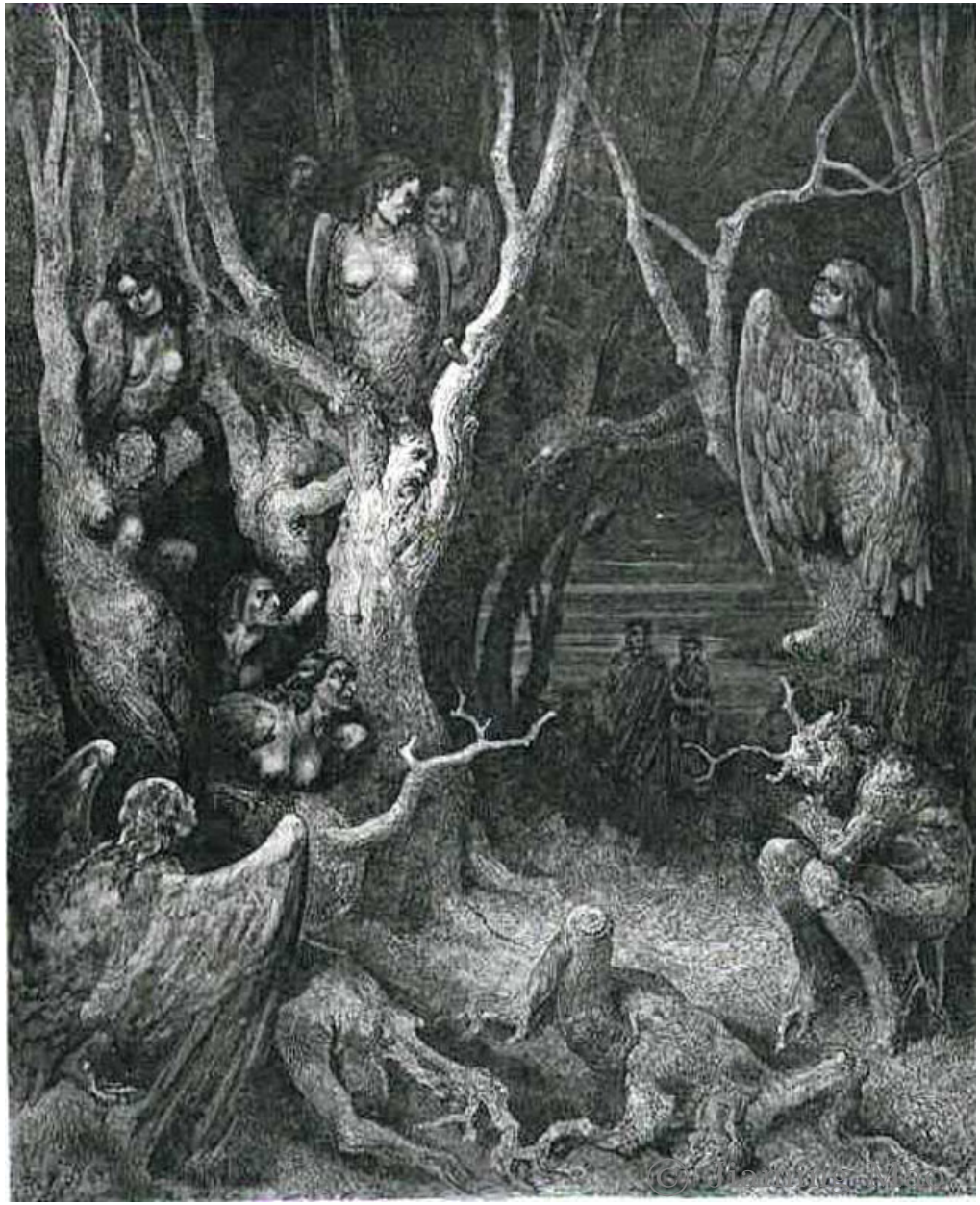
Road Map

... and journeys to the *realm of patterns*, where patterns in data are discovered and used to make predictions,



Road Map

... along the way he encounters the false prophets of *correlation equals causation*,



Road Map

... and then arrives at the *land of causality*, where people are serious about whether any two sets of observed phenomena are linked causally,

Road Map

... from where our hero finally reaches the *mount of scientific discovery*, where the mechanisms that generate the observed phenomena are investigated in the hope of attaining true knowledge about the world.

Statistical Learning

- Given variables x and y, how do we characterize the statistical relationship between the two?
  - $p(x, y)$: joint distribution of x and y[1]
- Oftentimes, we may not be interested in characterizing the full joint distribution $p(x, y)$. Instead, we are interested in predicting the value of y based on observed x.
  - We want to find a function $f(x)$ for predicting y given values of x.

[1] In this lecture, we use $p(x)$ to both denote the probability mass function (pmf) if x is a discrete random variable and the probability density function (pdf) if x is a continuous random variable.

Statistical Learning

Let

$$y = f(x) + e$$

, where e is an error term.

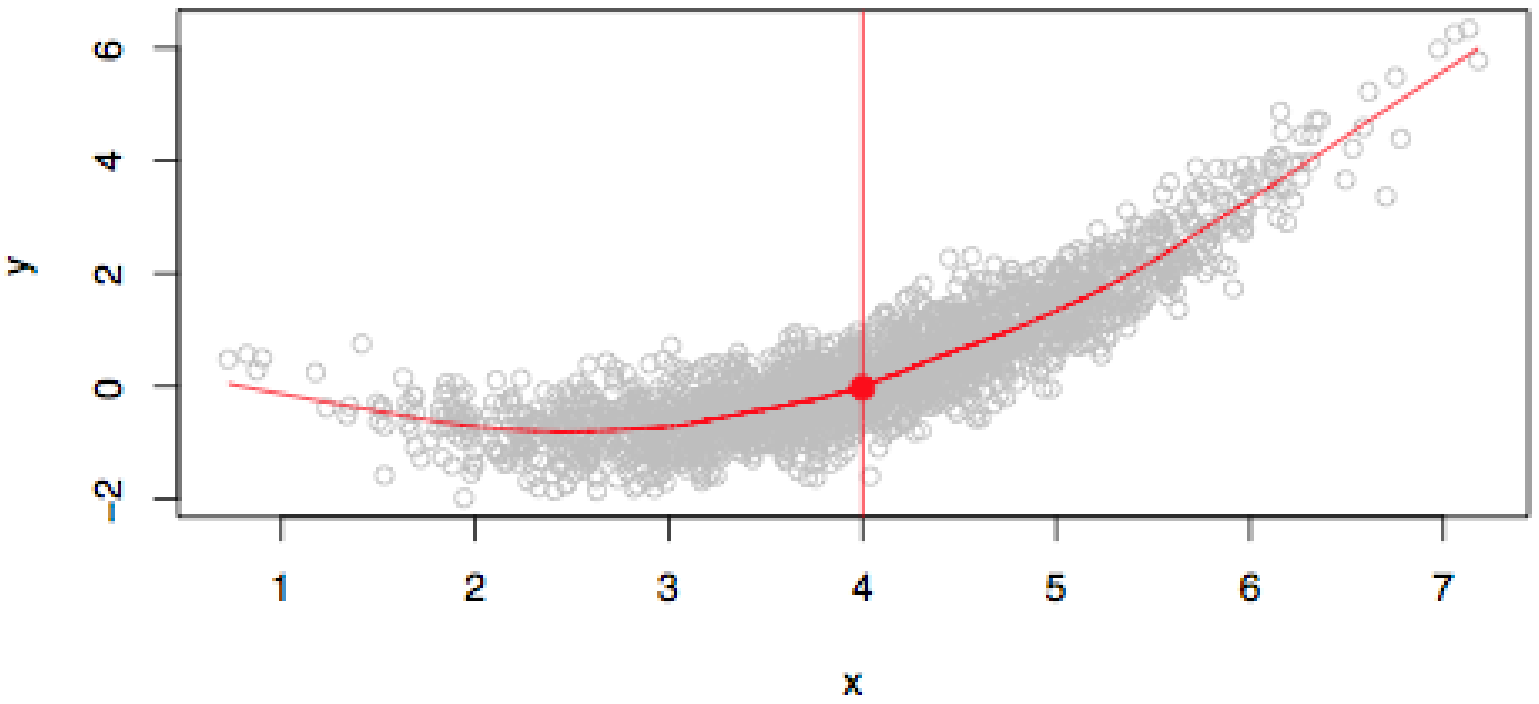
What is the function f that produces the best prediction of y given x?

- Depends on how we measure "best." Common choice: minimizing the expected squared-error loss[2] $\mathbb{E}\left[(y - f(x))^2\right] \Rightarrow f(x) = \mathbb{E}[y|x]$.
- $f(x) = \mathbb{E}[y|x]$ is the **target function** that we want to learn[3].

[2] Also commonly called the **mean squared error (MSE)**.

[3] **Learning** is also called **estimation**. We will use the two terms interchangeably.

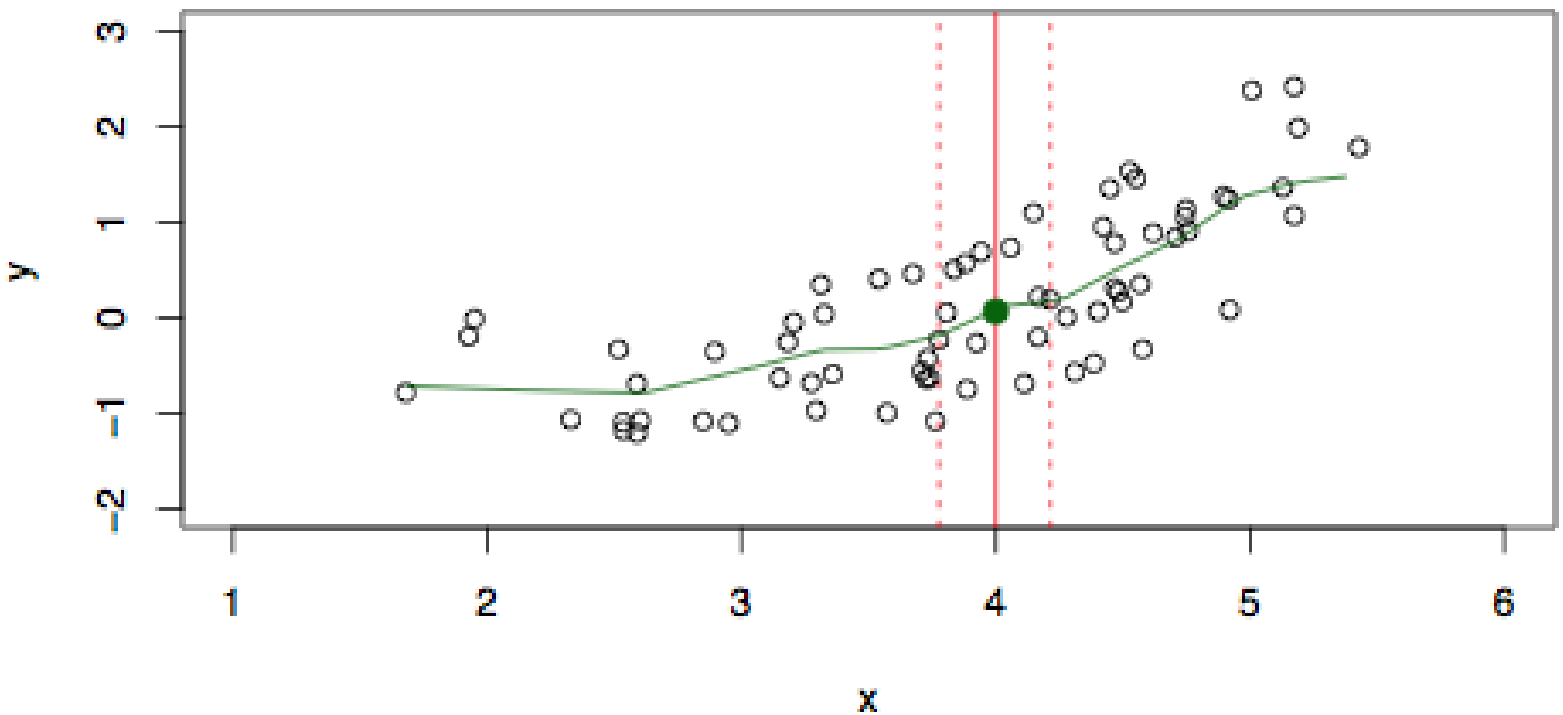
Learning f



$$\widehat{f}\,(x=4)=Ave\,(y|x=4)$$

Learning f

- Typically we have few if any data points at a specific value of x.
- One solution: relax the set of x over which y is averaged.



$$\widehat{f}\,(x=4) = Ave\,(y|x \in \mathcal{N}\,(x=4))$$

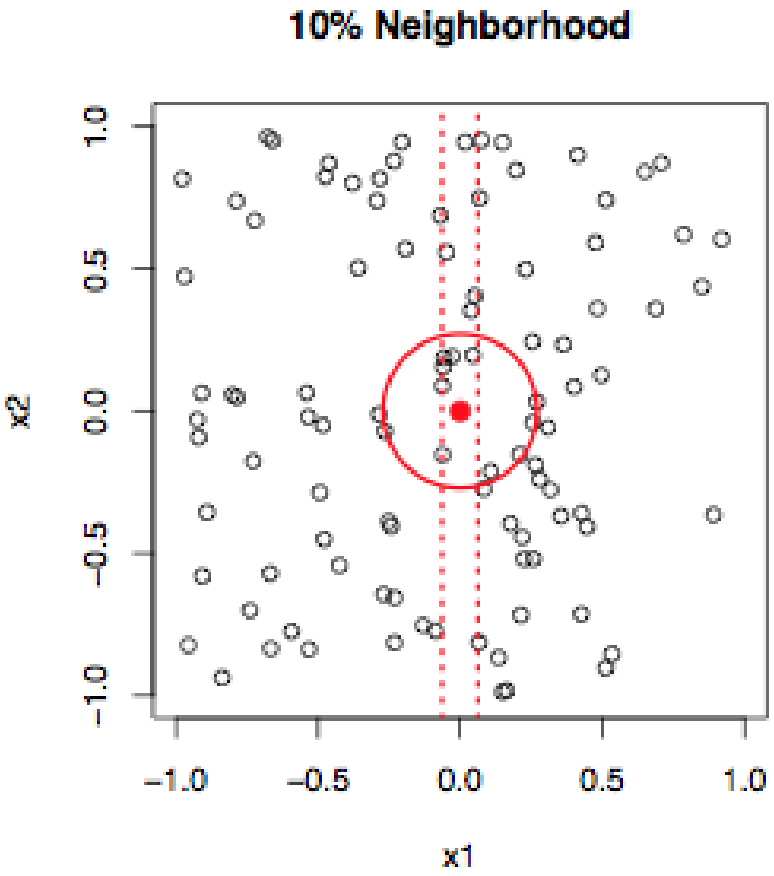
, where $\mathcal{N}\,(x)$ is some neighborhood of x.

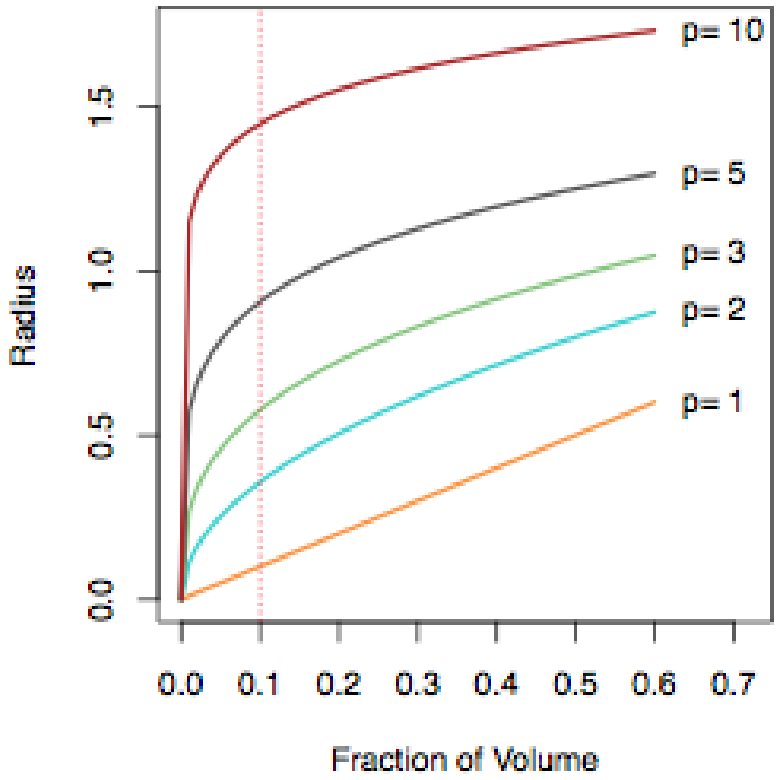
Learning f

- When x is multi-dimensional, i.e. $x = (x_1, \ldots, x_p)$, nearest neighbor averaging can work well for small p and large N[4].
- Nearest neighbor methods can be lousy when p is large, because neighbors tend to be far away in high dimensions.
  - This is called the **curse of dimensionality.**

[4] N : the number of data points

Learning f





Nearest neighbor and the curse of dimensionality

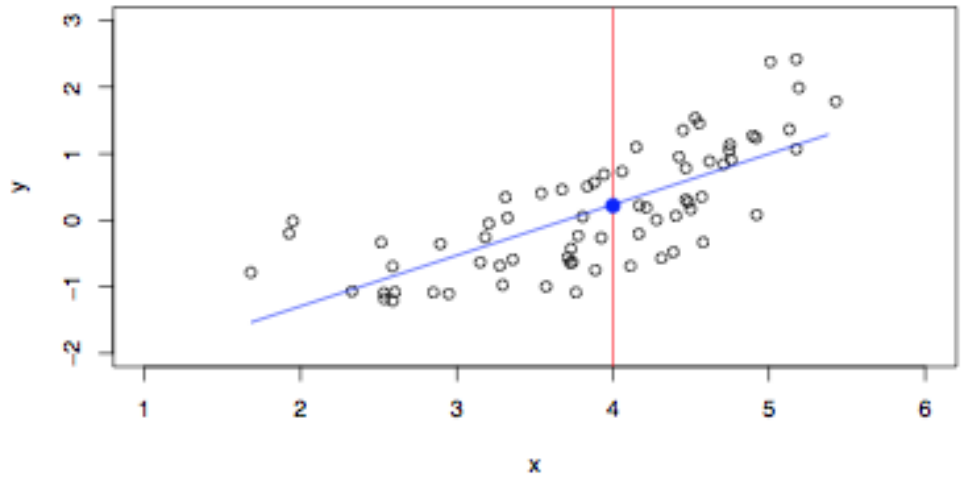
Learning f

- **Parametric methods**[5] of estimating $f(x)$ assume a specific functional form with a fixed number of parameters.
  - Linear regression: $f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \beta' x$
- **Nonparametric methods** do not make explicit assumptions about the functional form of $f(x)$[6].
  - Nearest neighbor averaging is a nonparametric method.

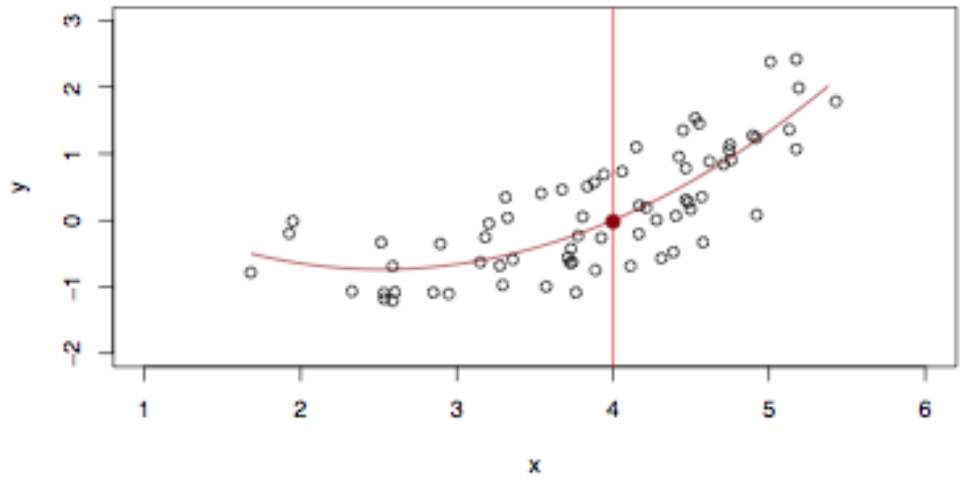
[5] We will use the terms "statistical method" and "statistical model" interchangeably.

[6] We will also learn methods that make some assumptions about the functional form of $f(x)$, but allow the number of parameters to grow with data.

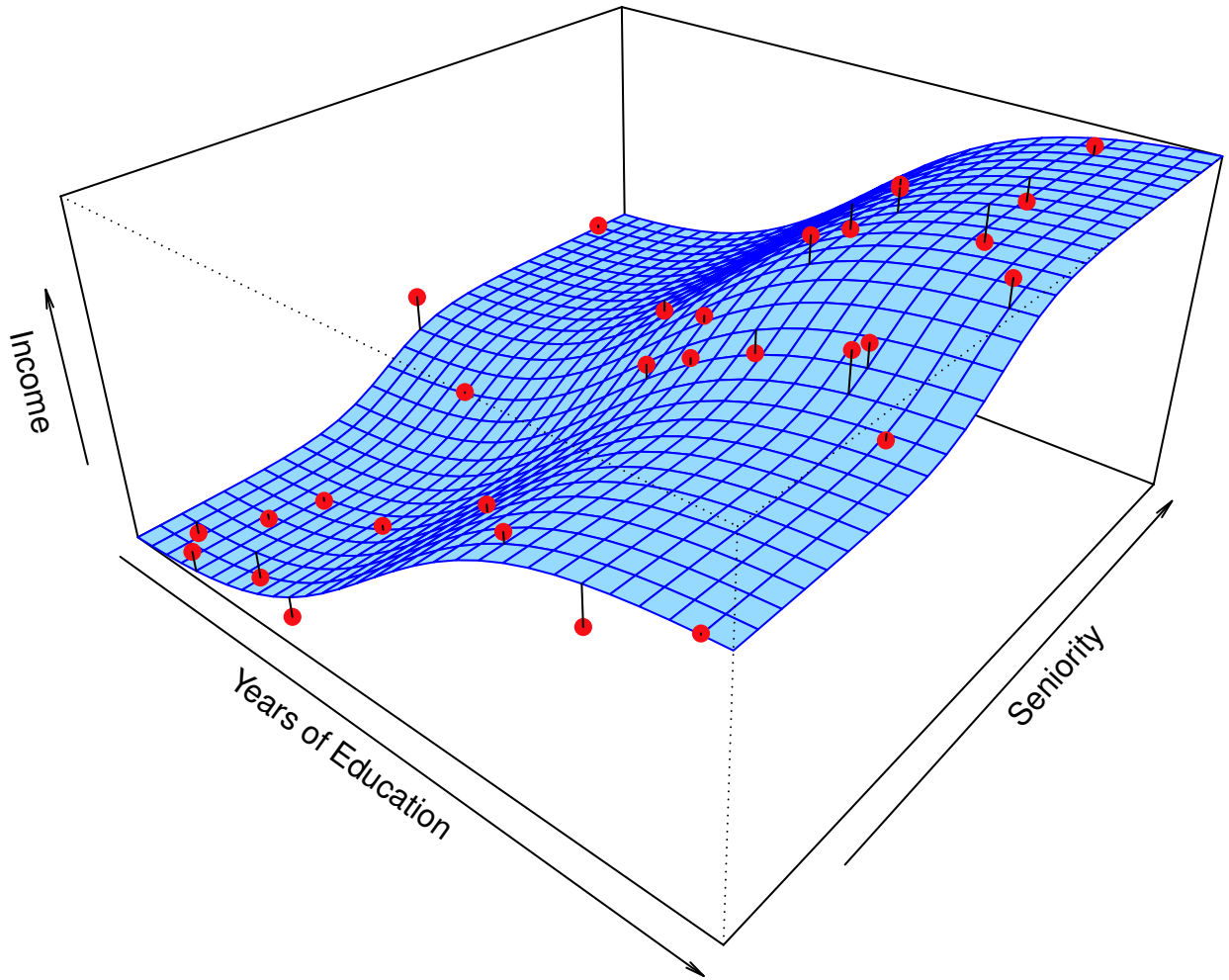
A linear model $\hat{f}_L(X) = \hat{\beta}_0 + \hat{\beta}_1 X$ gives a reasonable fit here



A quadratic model $\hat{f}_Q(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$ fits slightly better.

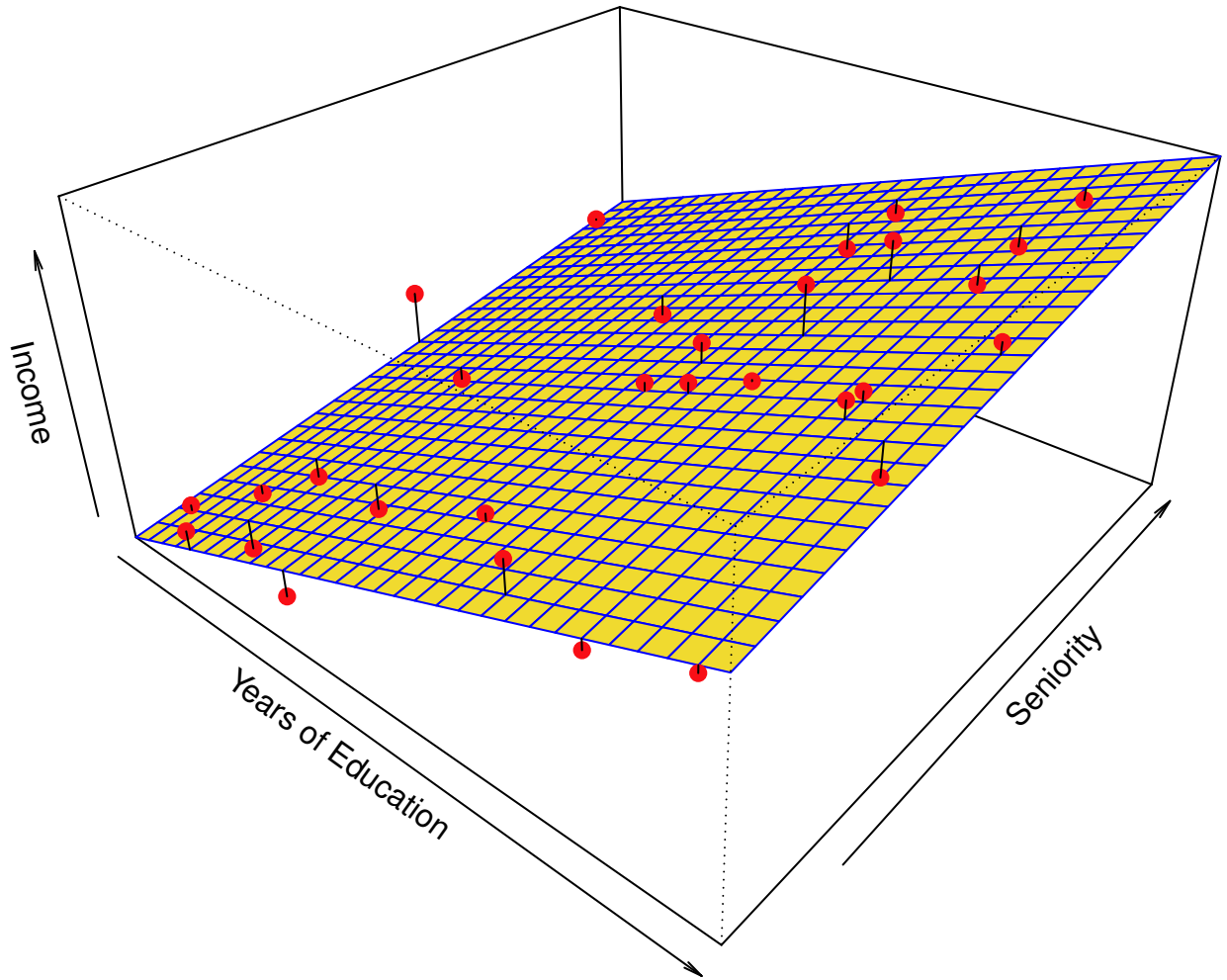


Learning f



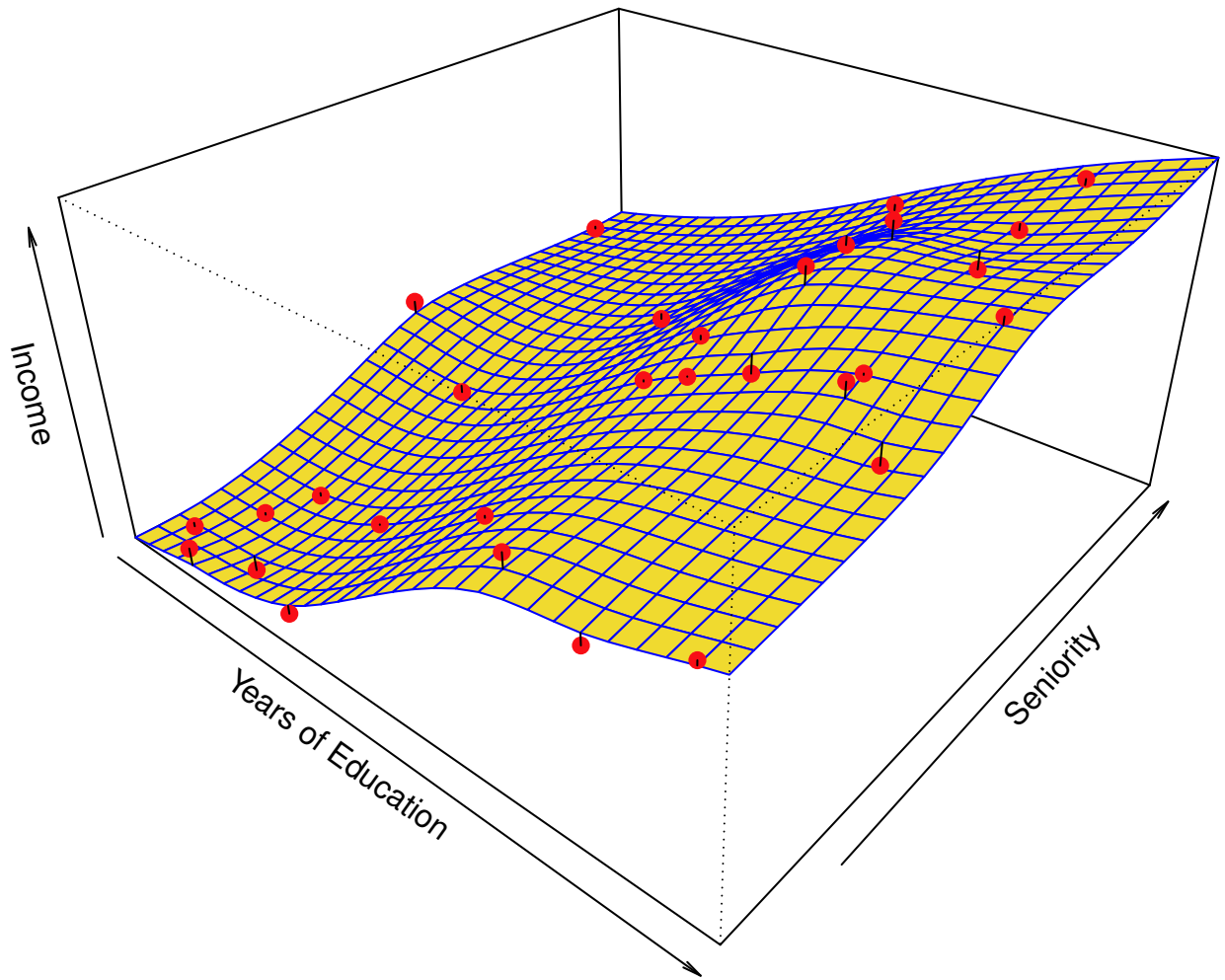
True f

Learning f



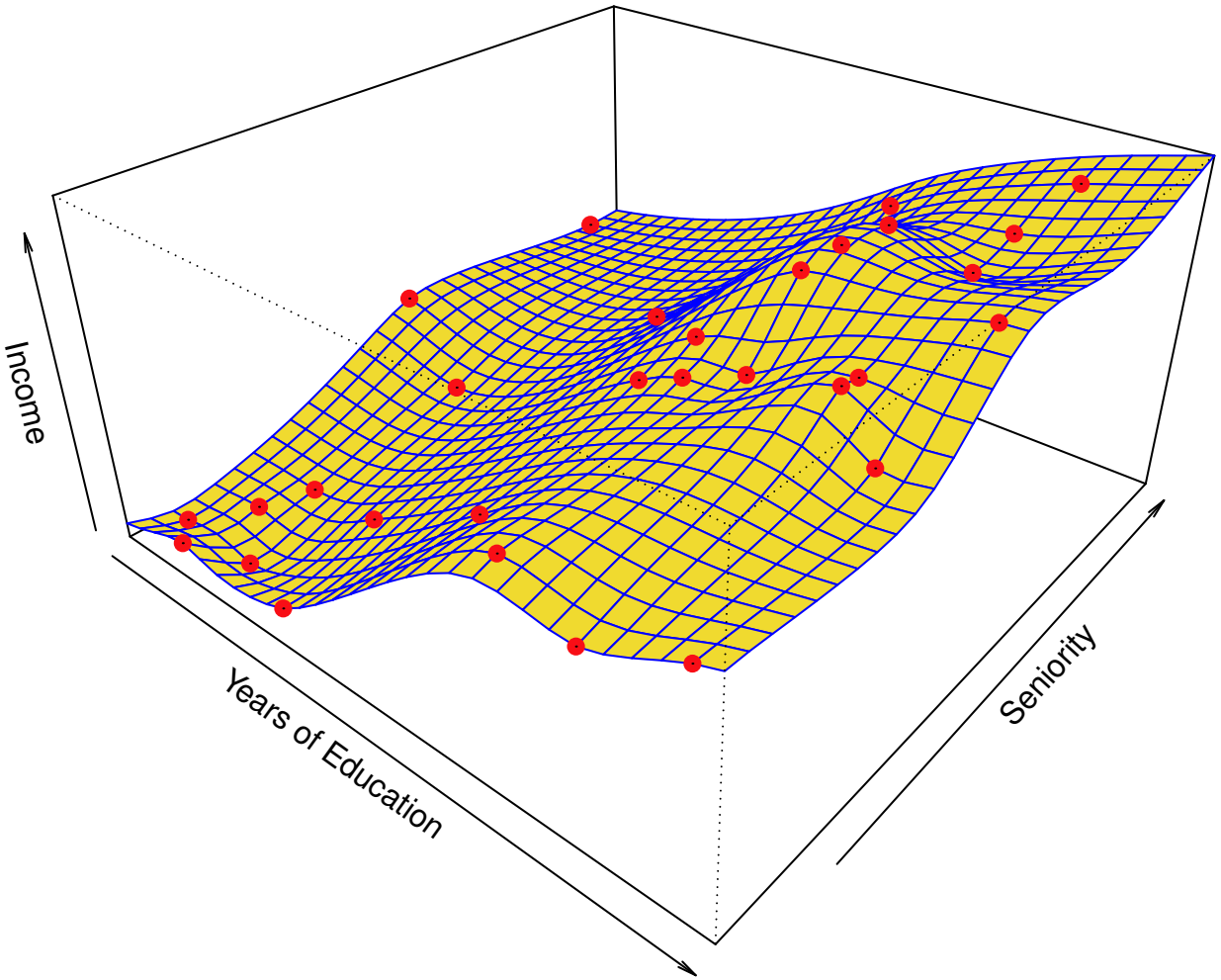
Linear Fit

Learning f



Thin-plate Spline Fit (Smooth)

Learning f



Thin-plate Spline Fit (Rough)

Here $\widehat{f}$ fits the data perfectly: $\widehat{f}(x)$ contains not only $f(x)$ but also e.

Assessing the Goodness of Fit

Let $\mathcal{D}_{TR} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ denote the data on which we estimate f. This is called **training data**.

We can assess how well $\widehat{f}$ fits the training data by calculating the **training error**:

$$\text{error}_{TR} = \frac{1}{N} \sum_{i \in \mathcal{D}_{TR}} \left(y_i - \widehat{f}(x_i) \right)^2$$

However, what we are really interested in is how well $\widehat{f}$ predicts previously unseen data.

Assessing the Goodness of Fit

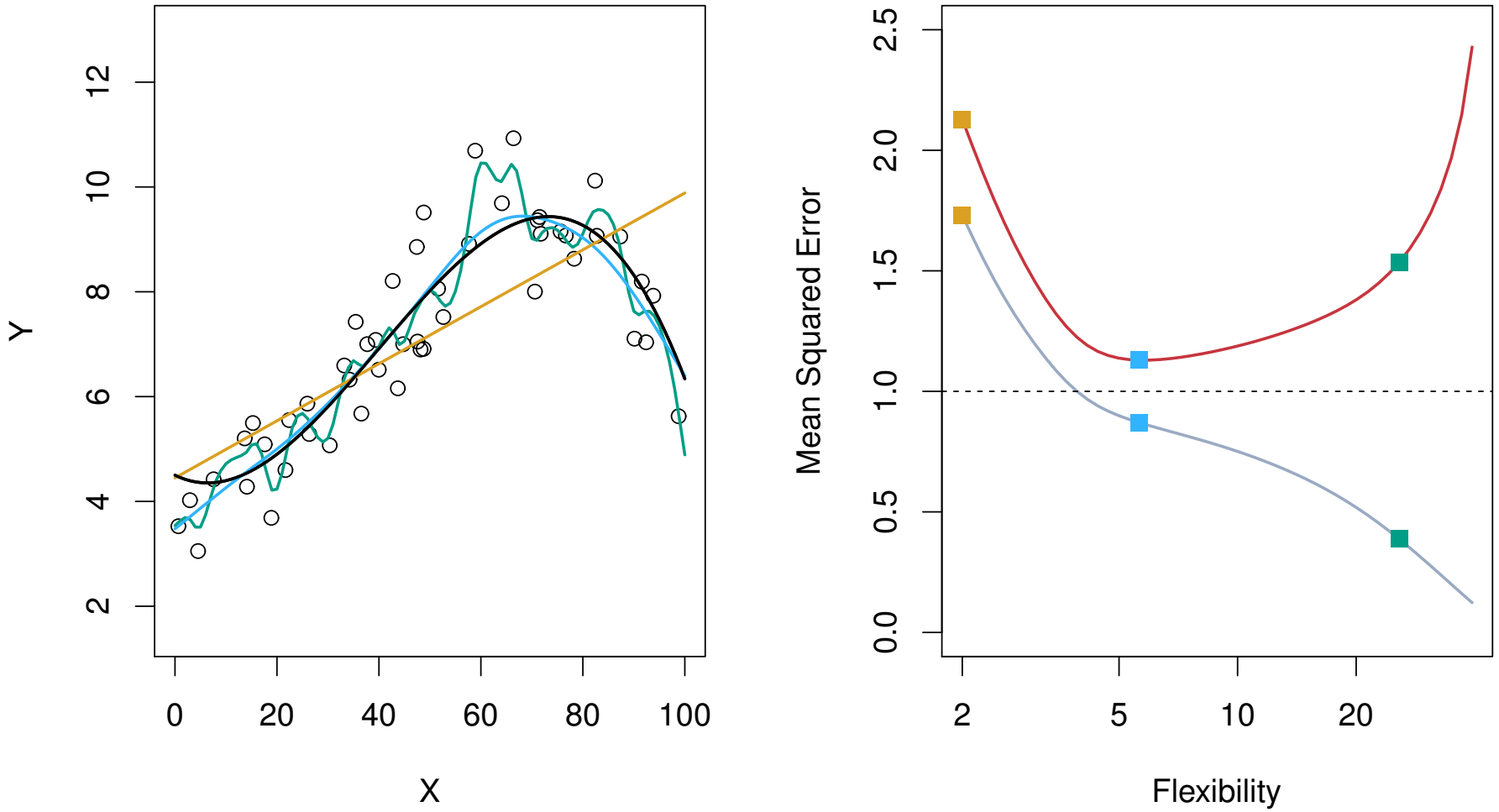
To this end, we can apply $\widehat{f}$ to a set of **test data**, $\mathcal{D}_{TE} = \{(x_1, y_1), \ldots, (x_M, y_M)\}$, and calculate the **test error**:

$$\text{error}_{TE} = \frac{1}{M} \sum_{i \in \mathcal{D}_{TE}} \left(y_i - \widehat{f}(x_i) \right)^2$$

When $M \to \infty$, $\text{error}_{TE} \to \mathbb{E}\left[\left(y - \widehat{f}(x) \right)^2 \right]$[7].

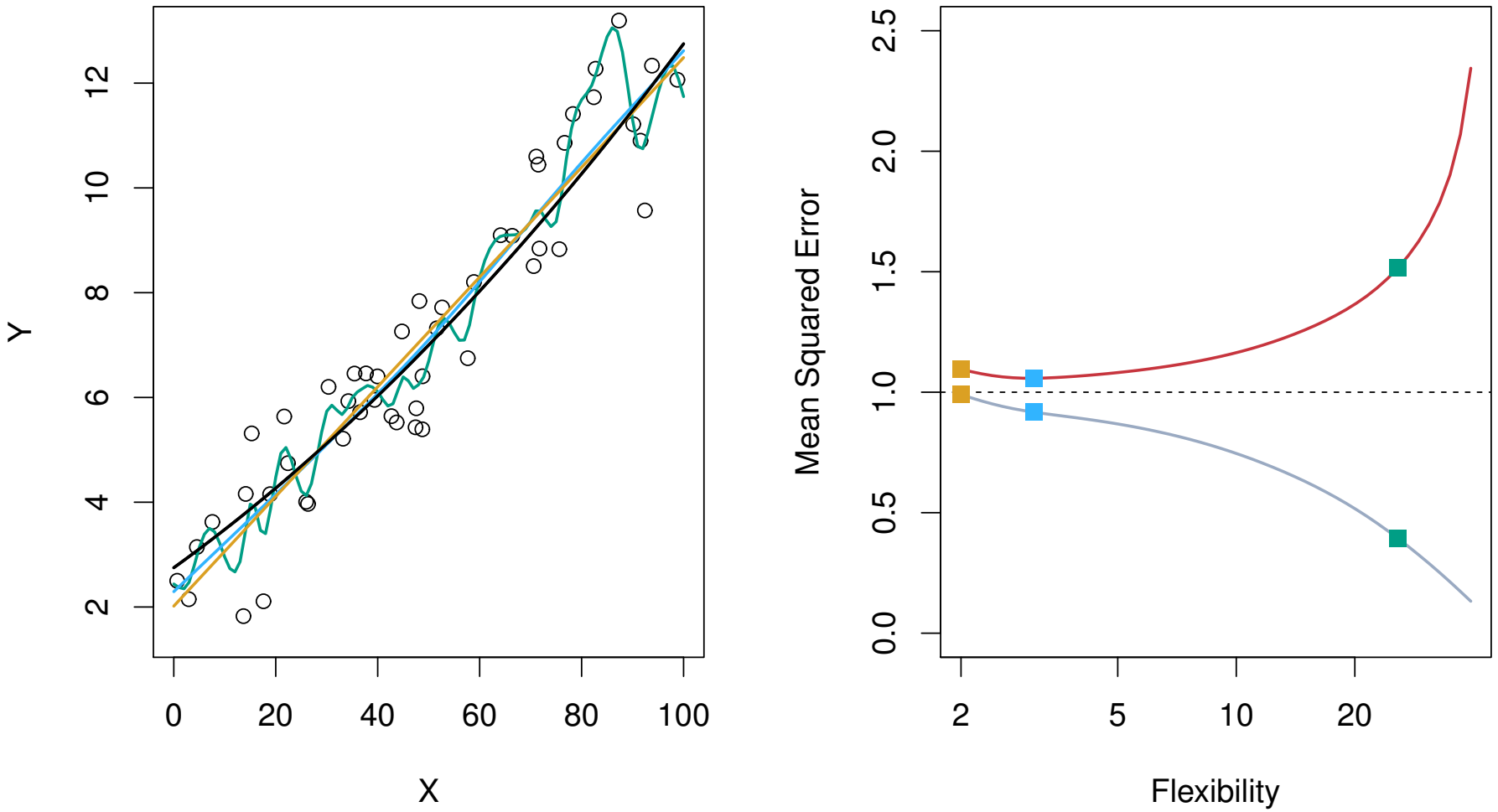
[7] $\mathbb{E}[(y - \widehat{f}(x))^2]$ is called the **expected error** or **prediction error**.

Assessing the Goodness of Fit



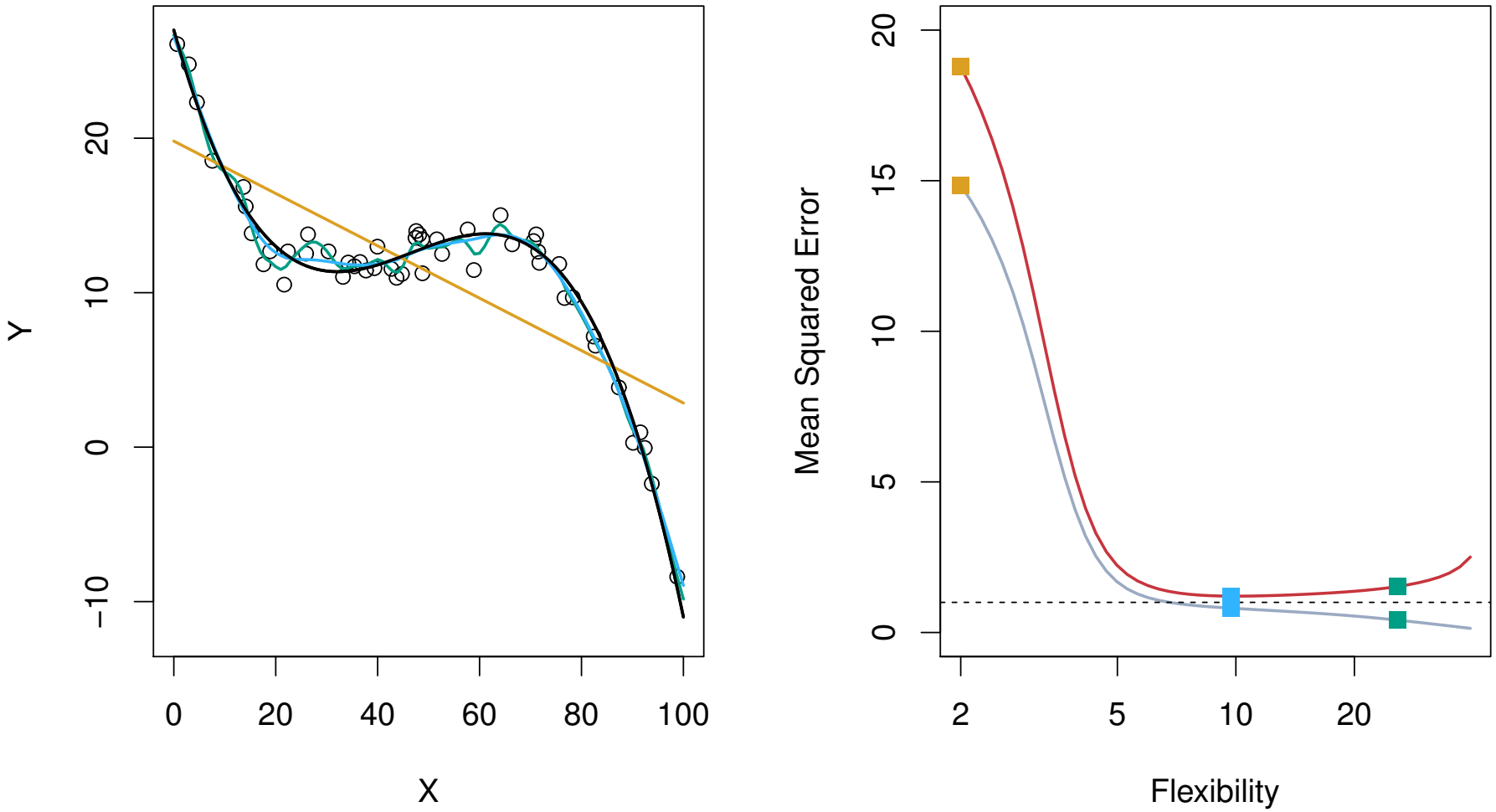
Left: true f (black), linear fit (orange), smoothing spline fits (blue & green).
Right: training error (grey), prediction error (red), Var(e) (dashed).

Assessing the Goodness of Fit



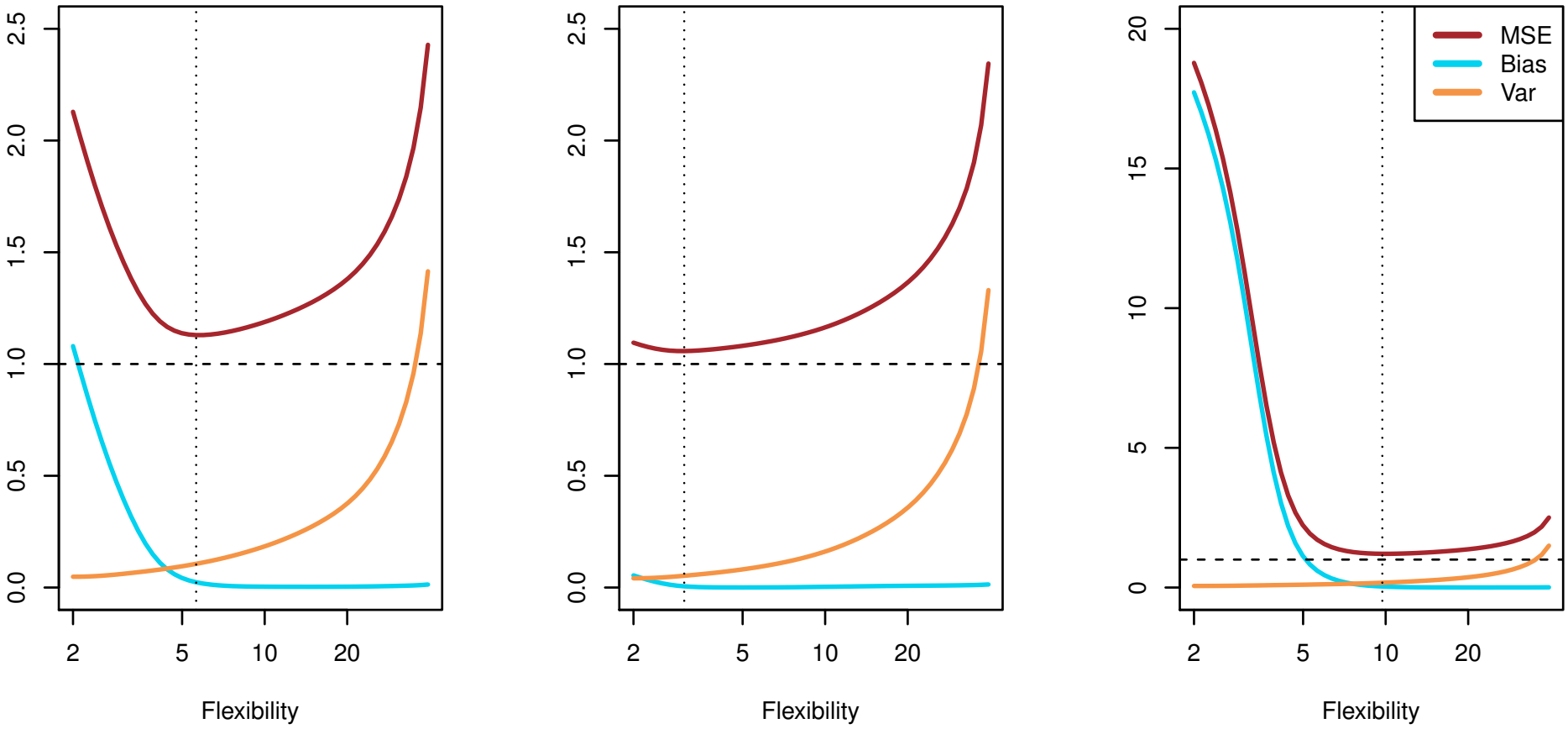
Left: true f (black), linear fit (orange), smoothing spline fits (blue & green).
Right: training error (grey), prediction error (red), Var(e) (dashed).

Assessing the Goodness of Fit



Left: true f (black), linear fit (orange), smoothing spline fits (blue & green).
Right: training error (grey), prediction error (red), Var(e) (dashed).

Assessing the Goodness of Fit



Bias-variance trade-off for the three examples

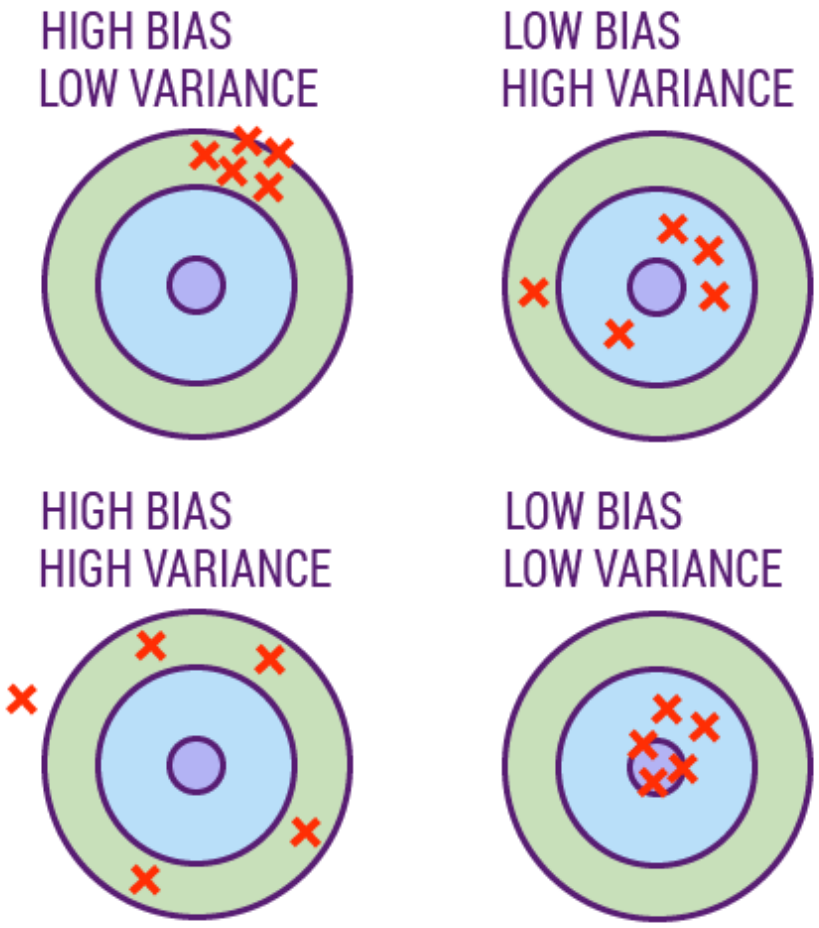
The Bias-Variance Trade-off

At a given x,

$$\mathbb{E}_{\mathcal{D}_{TR}}\left[\left(f\left(x\right)-\widehat{f}\left(x\right)\right)^2\right]=Var\left(\widehat{f}\left(x\right)\right)+\left(\text{bias}\left(\widehat{f}\left(x\right)\right)\right)^2$$

, where $\text{bias}\left(\widehat{f}\left(x\right)\right)\equiv\mathbb{E}_{\mathcal{D}_{TR}}\left[\widehat{f}\left(x\right)\right]-f\left(x\right)$.

The Bias-Variance Trade-off



The Bias-Variance Trade-off

- Intuitively, the bias term arises due to our model not able to capture the true f.
- The variance term arises because we have *limited* data.
  - $Var\left(\widehat{f}\right)$ refers to the amount by which $\widehat{f}$ would change if we estimate it using a different training data set.
  - $Var\left(\widehat{f}\right) = 0$ if we have access to the entire population.

The Bias-Variance Trade-off

- As a general rule, as model flexibility increases, bias $\left(\widehat{f}\right)$ will decrease and $Var\left(\widehat{f}\right)$ will increase.
  - More flexible models tend to have higher variance because they have the capacity to follow the data more closely. Thus changing any of the data points may cause the estimate $\widehat{f}$ to change considerably.

The Bias-Variance Trade-off

- As the flexibility of the model increases, we observe a monotone decrease in training error and a U-shape in prediction error.

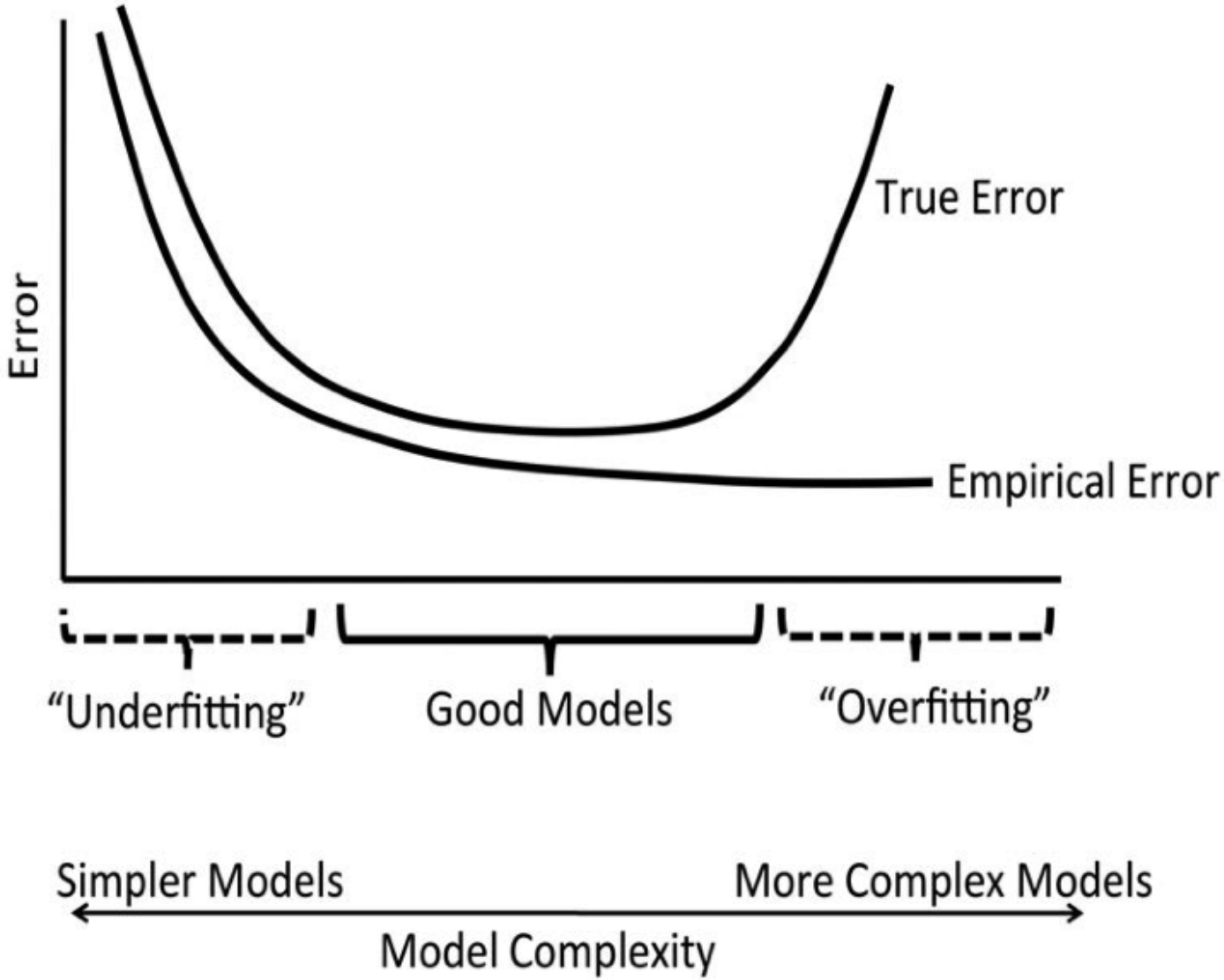
- This is due to the **bias-variance trade-off**: as model flexibility increases, the bias tends to initially decrease faster than the variance increases. Then at some point increasing flexibility has little impact on the bias but starts to significantly increase the variance.

The Bias-Variance Trade-off

- The bias-variance trade-off is a trade-off because it is easy to have a model with low bias but high variance (e.g., neighborhood averaging) or one with low variance but high bias (e.g., a constant model). The challenge lies in finding a model for which both the variance and the bias are low.

- **Overfitting** refers to the case in which a less flexible model would have yielded a smaller prediction error.

The Bias-Variance Trade-off

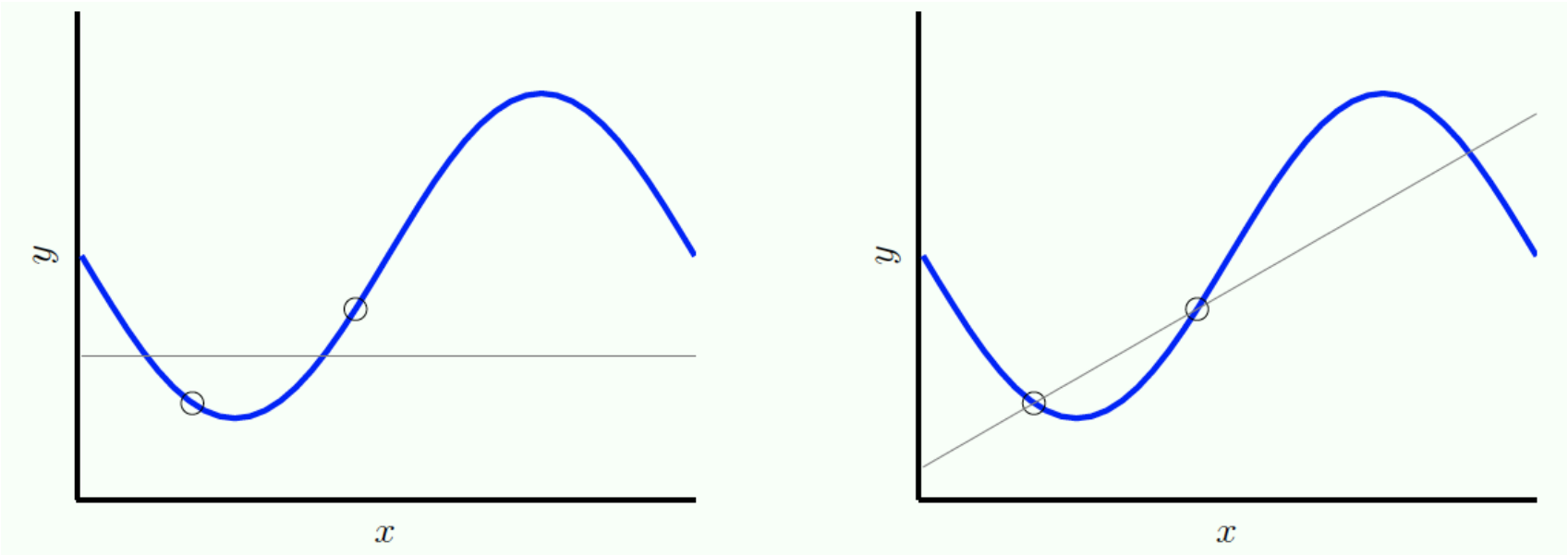


Choosing the Optimal Model

$$y = f(x) = \sin(\pi x)$$

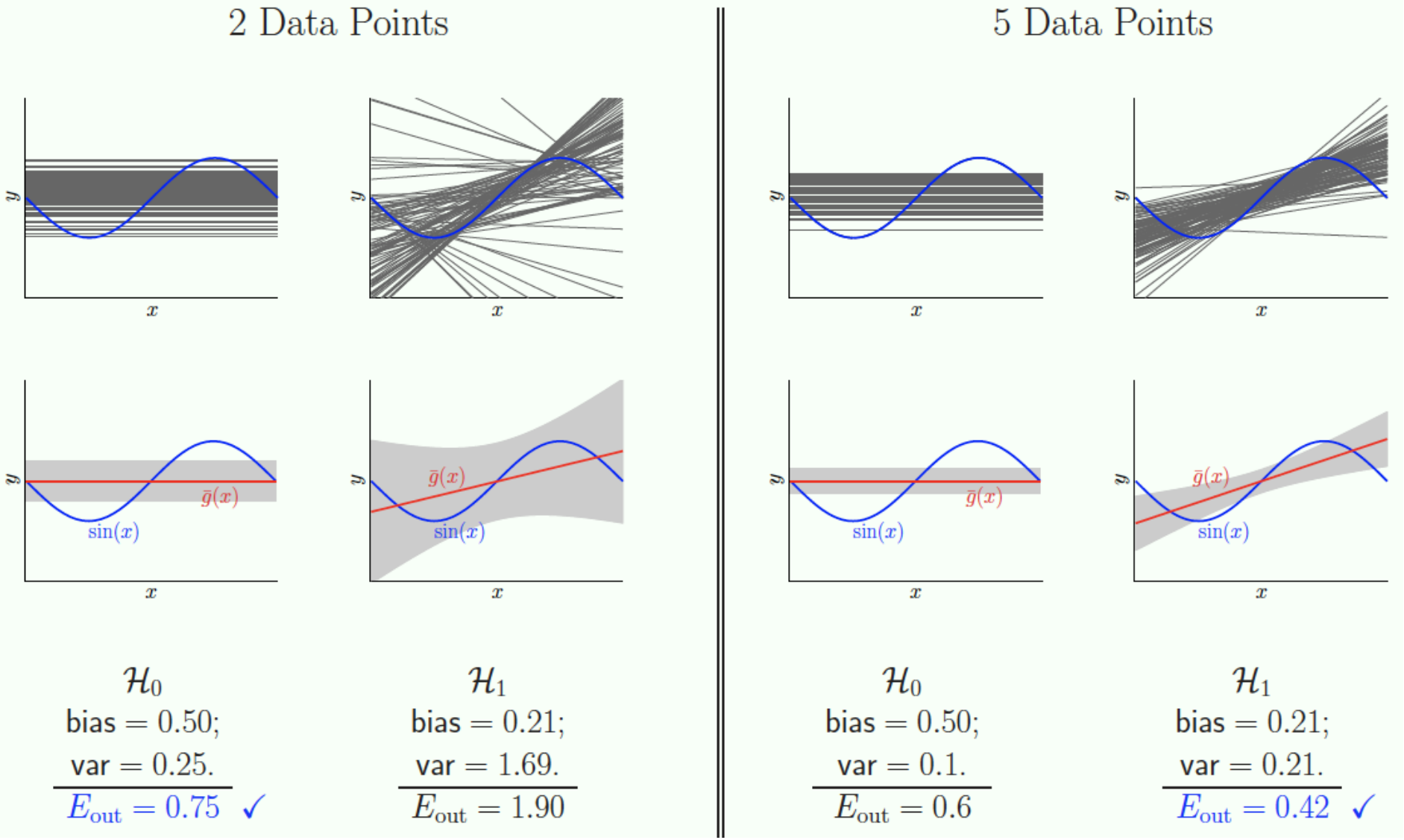
- Two models:

$$\mathcal{H}_0 : h(x) = b$$
$$\mathcal{H}_1 : h(x) = ax + b$$



2 data points

Choosing the Optimal Model



Choosing the Optimal Model

Optimal model complexity depends on:

1. Complexity of the true f
2. Sample size

Adaptive Statistical Models

- Modern machine learning methods can be characterized as **adaptive statistical models** that adaptively choose their complexity based on the data.
  - The lasso with $p-$dimensional features is a constant model at its simplest and a $p-$dimensional linear regression model at its most complex.
  - A decision tree at its simplest is a (piece-wise) constant model while at its most complex is a nonparametric neighbor averaging method.
  - A neural network becomes a linear model when its weights approach zero, but can increase its complexity to approximate any functional form.

Adaptive Statistical Models

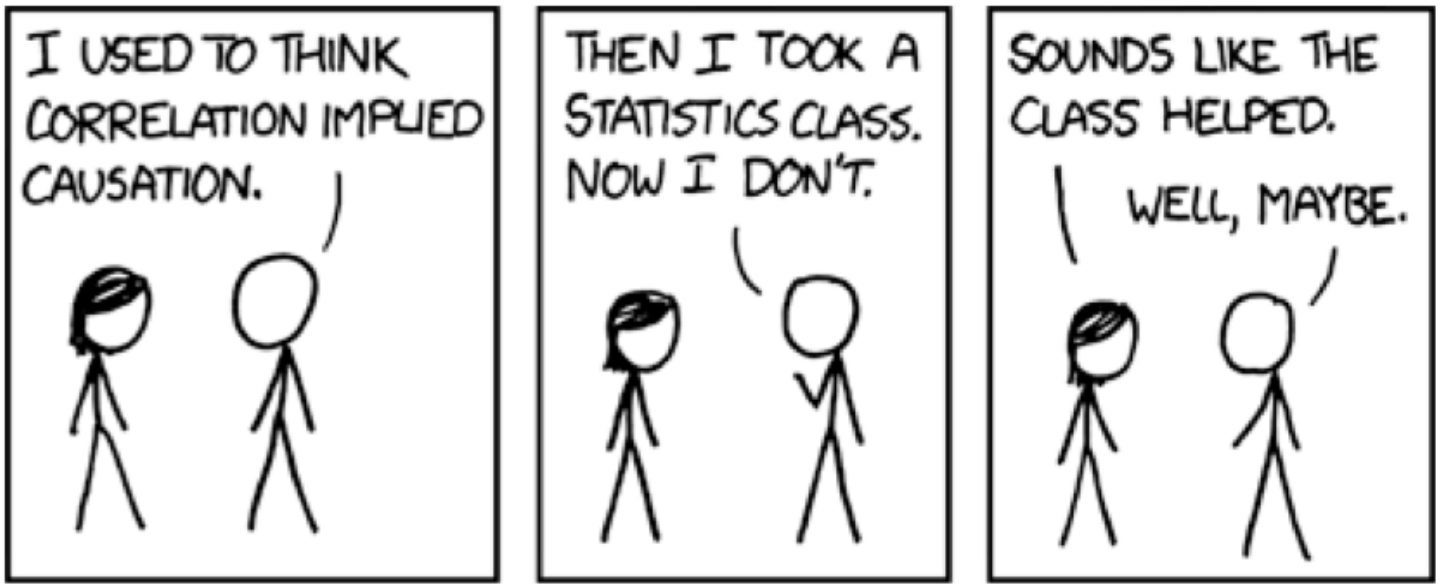
"In this paper, we review and apply several popular methods from the machine learning literature to the problem of demand estimation ... we compare these methods to standard econometric models that are used by practitioners to study demand ... we used sales data on salty snacks [from] scanner panel data from grocery stores ... In our results, we find that the six models we use from the statistics and computer science literature predict demand out of sample in standard metrics much more accurately than a panel data or logistic model." – Bajari et al. (2015)

Adaptive Statistical Models

	Validation		Out-of-Sample		
	RMSE	Std. Err.	RMSE	Std. Err.	Weight
Linear	1.169	0.022	1.193	0.020	6.62%
Stepwise	0.983	0.012	1.004	0.011	12.13%
Forward Stagewise	0.988	0.013	1.003	0.012	0.00%
Lasso	1.178	0.017	1.222	0.012	0.00%
Random Forest	0.943	0.017	0.965	0.015	65.56%
SVM	1.046	0.024	1.068	0.018	15.69%
Bagging	1.355	0.030	1.321	0.025	0.00%
Logit	1.190	0.020	1.234	0.018	0.00%
Combined	0.924		0.946		100.00%
# of Obs	226,952		376,980		
Total Obs	1,510,563				
% of Total	15.0%		25.0%		

Bajari et al. (2015)

Causal Inference



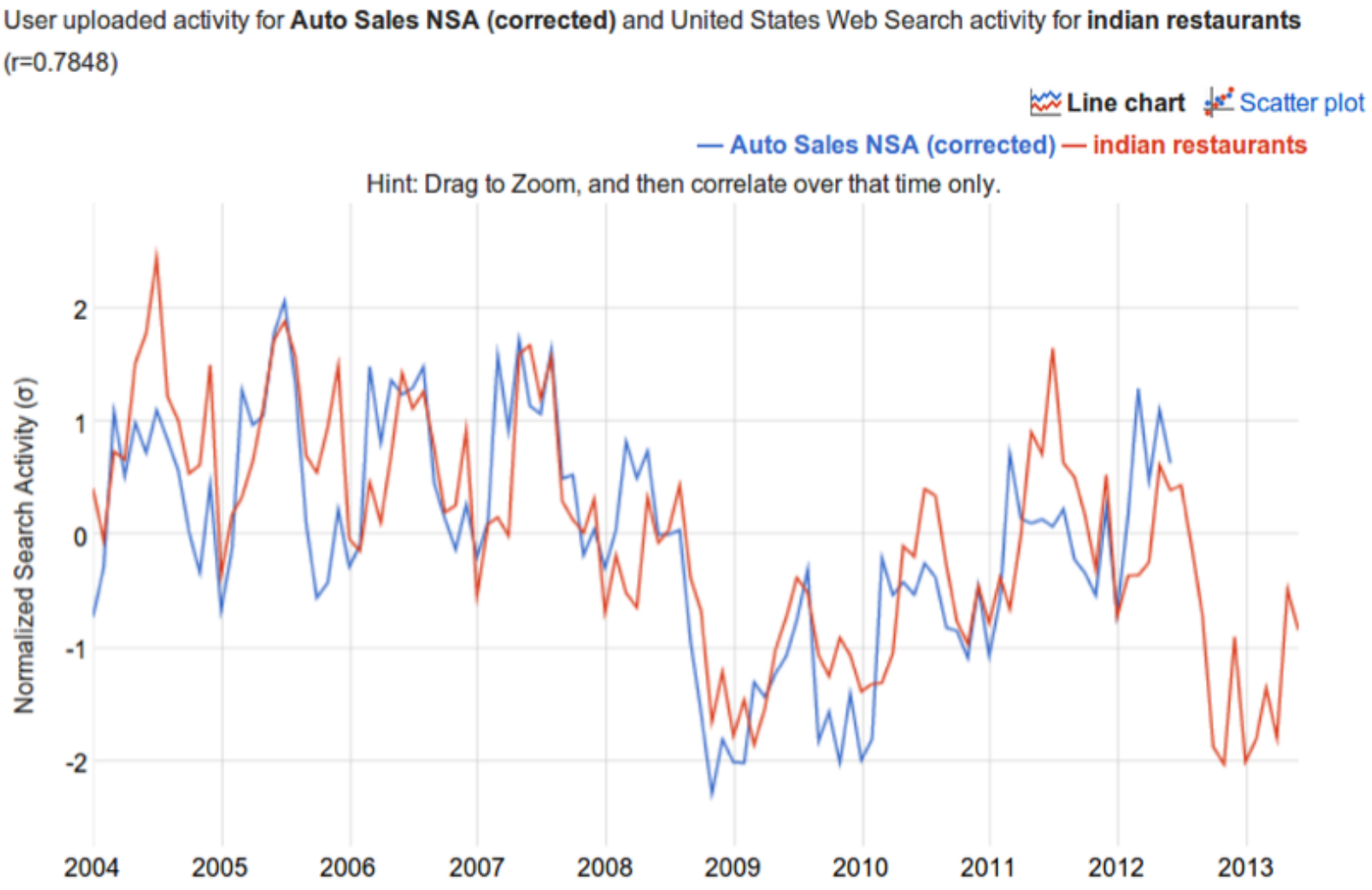
Causal Inference

- Learning the statistical relationship between x and y tells us nothing about whether there exists a causal relationship between them.

- **Causal inference** is concerned with the following questions:

  1. Does x have a causal effect on y? If so, how large is the effect? (**causal effect learning**)
  2. If a causal effect exists, what is the mechanism by which it occurs? (**causal mechanism learning**)

Correlation does not imply Causation



Automobile Sales and Search for Indian Restaurants

Seeing vs. Doing

The do operator:

$$\text{do}(x = a) : \text{set } x = a$$

- Barometer readings are useful for predicting rain:

$$\Pr(\text{rain} \mid \text{barometer} = \text{low}) > \Pr(\text{rain} \mid \text{barometer} = \text{high})$$

- But hacking a barometer won't change the probability of raining:

$$\Pr(\text{rain} \mid \text{do}(\text{barometer} = \text{low})) = \Pr(\text{rain} \mid \text{do}(\text{barometer} = \text{high}))$$

Seeing vs. Doing

- Doing: if x has a causal effect on y, then we can change x and expect it to cause a change in y.
- Seeing: If x is correlated[8] with y but does not have a causal effect on y, then we can only observe the correlation without the ability to change y by manipulating x.

[8] We use the term "correlation" in its broad sense to mean statistical dependence (association).

Causal vs. Statistical Predictions

- **Causal prediction**: What will y be if I *set* $x = a$?
  - $\mathbb{E}[y|\text{do}(x = a)]$[9]
- **Statistical prediction**: What will y be if I *observe* $x = a$?
  - $\mathbb{E}[y|x = a]$

[9] Assuming we minimize the expected L2 loss in prediction.

Causal Effect Learning

- To learn $f(x) = \mathbb{E}[y|\text{do}(x)]$, the simplest way is to "just do it".
- Let a be a possible value of x. Randomly select individual units, *set* their $x = a$, and observe the resulting y. In this way, we can *generate data* from $p(y|\text{do}(x))$.
  - This is in essence what a randomized experiment does.
- A nonparametric estimator for $f(x)$ is then

$$\widehat{f}(x = a) = Ave(y|x = a)$$

Causal Effect Learning

Randomized Experiment

- Consider $x \in \{0, 1\}$. Suppose we are interested in learning the causal effect of $x = 1$ on y.

- Given a set of experimental units, a **randomized controlled trial** (**RCT**) randomly selects a subset of individual units – call them the **treatment group** – to receive $x = 1$, and assign $x = 0$ to the rest of the experimental units – called them the **control group**.

Randomized Experiment

Using the experimental language, x is called **treatment** and y is called **outcome**. The **average treatment effect** (**ATE**)[10] is defined as

$$\begin{aligned} \text{ATE} &\doteq \mathbb{E}\left[y|\text{do}\left(x=1\right)\right] - \mathbb{E}\left[y|\text{do}\left(x=0\right)\right] \\ &\overset{[1]}{=} \mathbb{E}\left[y|x=1\right] - \mathbb{E}\left[y|x=0\right] \end{aligned}$$

, where [1] follows because randomized experiments generate data from $p\left(y|\text{do}\left(x\right)\right)$, therefore $\mathbb{E}\left[y|x\right] = \mathbb{E}\left[y|\text{do}\left(x\right)\right]$.

For data generated by randomized experiments, correlation implies causation.

[10] The terms "treatment effect" and "causal effect" are used interchangeably.

Randomized Experiment

The
Design of Experiments

By

Sir Ronald A. Fisher, Sc.D., F.R.S.

Honorary Research Fellow, Division of Mathematical Statistics, C.S.I.R.O., University of Adelaide; Foreign Associate, United States National Academy of Sciences; and Foreign Honorary Member, American Academy of Arts and Sciences; Foreign Member of the Swedish Royal Academy of Sciences, and the Royal Danish Academy of Sciences and Letters; Member of the Pontifical Academy; Member of the German Academy of Sciences (Leopoldina); formerly Galton Professor, University of London, and Arthur Balfour Professor of Genetics, University of Cambridge.

HAFNER PRESS
A Division of Macmillan Publishing Co., Inc.
New York
COLLIER MACMILLAN PUBLISHERS
London

The Experimental Ideal and Its Limitations

- For many causal inference problems, RCTs are impossible or impractical to run.
  - infeasibility (e.g., monetary policy)
  - ethical reasons (e.g., smoking and lung cancer)
  - cost and duration (e.g., childhood intervention and adult outcomes)

The Experimental Ideal and Its Limitations

- Results from many RCT studies suffer from a lack of **external validity** or **inability to scale**.
  - The ATE computed from an RCT study represents the average treatment effect in the *experiment population*, which is often different from – and significantly smaller than – the *target population*[11].
  - A treatment may have very different effects when it is applied to a small RCT sample and when it is applied to a significant proportion of a large population due to equilibrium effects.

[11] Fundamentally, this problem is due to the highly heterogeneous nature of many treatment effects.

External Validity

"Psychology is the study of psychology students." – Anonymous

Vol 466|1 July 2010 nature

OPINION

Most people are not WEIRD

To understand human psychology, behavioural scientists must stop doing most of their experiments on Westerners, argue **Joseph Henrich**, **Steven J. Heine** and **Ara Norenzayan**.

A 2008 survey of the top psychology journals found that 96% of subjects were from Western, educated, industrialized, rich and democratic (WEIRD) societies – particularly American undergraduates.

Observational Studies

For observational data, correlation no longer implies causation.

Consider the following example: suppose we observe patients at two hospitals:

Hospital	Sample Size	Recovery Rate
A	1274	97%
B	569	72%

Observational Studies

Based on this observation, can we conclude that hospital A is better?

- If patients are randomly administered to hospitals – in other words, if the data come from a randomized experiment, then yes.
- In observational studies, however, it could well be that hospital B is associated with worse outcomes because it is actually better, so that people with worse health problems *choose* to visit B.

Observational Studies

- In this case, let x denote hospital choice and y denote recovery rate. Then

$$\mathbb{E}[y|x] \neq \mathbb{E}[y|\text{do}(x)]$$

: when we observe a person visiting hospital B, we would expect a lower recovery rate ($\mathbb{E}[y|x = B]$) – because she is likely sicker – than the recovery rate we would expect if we randomly assign a person to hospital B ($\mathbb{E}[y|\text{do}(x = B)]$).

- This is called **self-selection effect** or **self-selection bias**.

Observational Studies

- Self-selection is of central concern to causal inference based on observed socio-economic data generated by individual choices.
- When individuals choose their own treatments, those who choose to receive a treatment can be *systematically* different from those who choose not to. If we compare their outcomes directly, then we are comparing apples with oranges[12].

[12] Note that such self-selection effect does not exist under random assignment of hospitals because the patients administered to each hospital would be similar. Comparing their outcomes would be comparing apples with apples.

Observational Studies

- To conduct valid causal inference on the effectiveness of hospital treatment, we need to compare recovery rates of patients with the same degree of illness who visit each hospital, i.e., we need compare apples with apples.

- Let z denote patient health prior to hospital visit, then

$$\mathbb{E}[y|\text{do}(x), z] = \mathbb{E}[y|x, z]$$

Conditional on patient illness, correlation between hospital choice and recovery rate implies causation[13]!

[13] To get the overall causal effect,

$$\mathbb{E}[y|\text{do}(x)] = \mathbb{E}_z[\mathbb{E}[y|\text{do}(x), z]] = \mathbb{E}_z[\mathbb{E}[y|x, z]]$$

Observational Studies

- As the example shows, to conduct causal inference on observational studies, we need to know how the data are generated (patients choose to visit different hospitals) and why outcomes may differ among treatment groups (patients administered to different hospitals are different in degree of illness and hospitals vary in their effectiveness – the latter is the treatment effect we are interested in)[14].
- Causal inference[15] requires an understanding of the causal mechanism that generates the data[16].

[14] But wait! What if hospitals accept different health insurance plans? Suppose hospital B accepts Medicare but hospital A does not, so that hospital B has many more older patients. How does this information change our causal inference?

[15] More precisely, causal effect learning. We will later discuss causal mechanism learning – how to discover the data-generating causal mechanism in the first place.

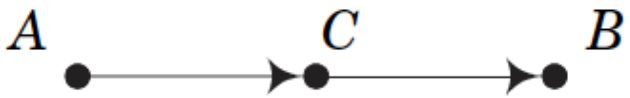
[16] As we will see, such understanding is not only necessary for observational studies but also necessary for interpreting and using experimental results. Without an understanding of – or making assumptions on – the underlying mechanism, any causal effect estimate is meaningless.

Causal Diagrams

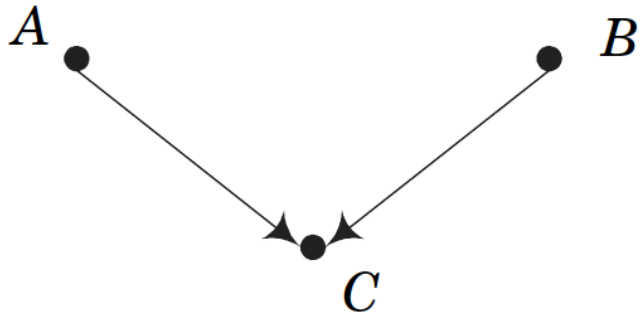
- **Causal diagrams** are graphs that represent causal relationships and therefore describe our qualitative knowledge about the causal mechanisms generating our observed data.

- In a causal diagram, the **nodes** (**vertices**) represent variables, with **directed edges** (**arrows**) representing direct causation. A sequence of connected nodes is called a **path**. The path is **causal** if all its arrows point in the same direction. Otherwise it is **noncausal**.

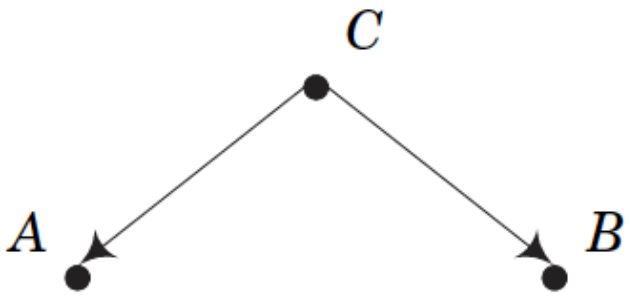
Causal Diagrams



(a) Mediation



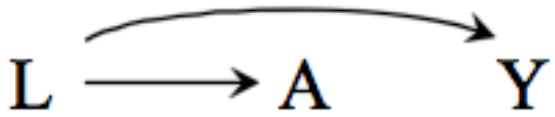
(c) Mutual causation



(b) Mutual dependence

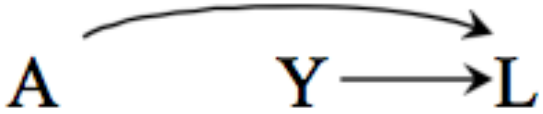
Basic patterns of causal relationships among three variables

Correlation and Causation



- L has a causal effect on both A and Y. A does not have a causal effect on Y. A depends on L and on *no other causes* of Y.
- L is called a **common cause** to A and Y.
- There exists an open path connecting A and Y: $A \leftarrow L \rightarrow Y$.
- A and Y are correlated: having information about A improves our ability to predict Y, even though A does not have a causal effect on Y.
- **Example:** A : carrying a lighter; Y : lung cancer; L : smoking

Correlation and Causation



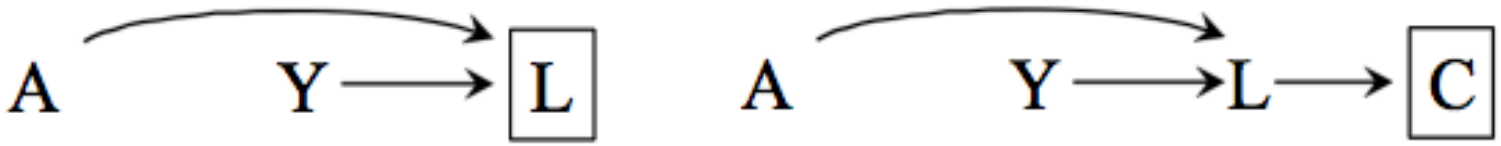
- Both A and Y have a causal effect on L. A does not have a causal effect on Y.
- L is called a **common effect** of A and Y.
- On the path $A \rightarrow L \leftarrow Y$, L is called a **collider**. The path is said to be blocked by the collider.
- A and Y are statistically independent.
- **Example:** A : family heart disease history; Y : smoking; L : heart disease

Correlation and Causation

Box indicates conditioning

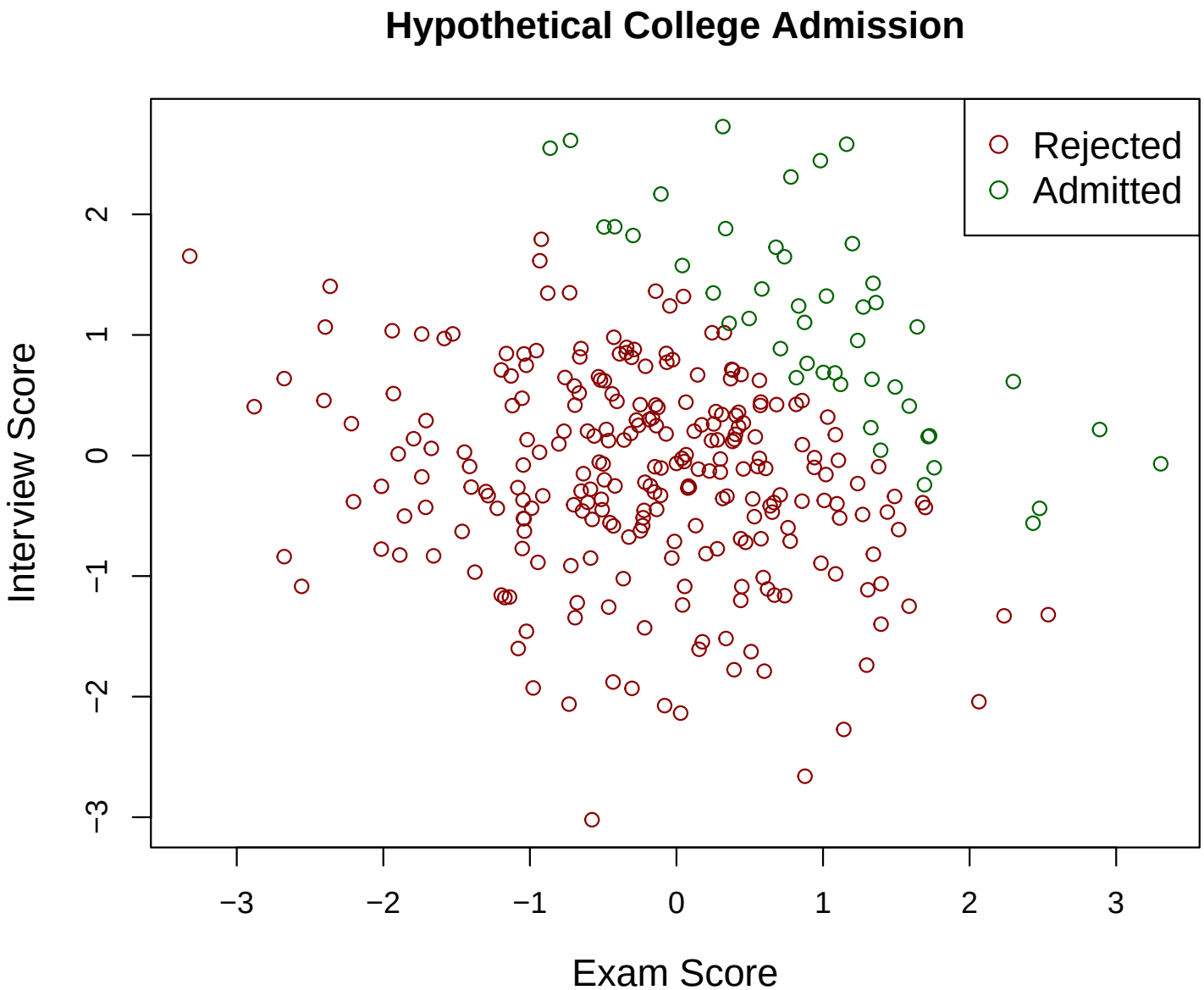
- A and Y are conditionally independent after conditioning on B and L, even though they are marginally correlated in both graphs.
- Conditioning on B and L block the paths $A \to B \to Y$ and $A \leftarrow L \to Y$.
- **Example:** (left) A : smoking; B : tar deposits in lung; Y : lung cancer

Correlation and Causation



- A and Y are conditionally correlated after conditioning on L and C, even though they are marginally independent.
- Conditioning on collider L or its descendent C opens the path $A \rightarrow L \leftarrow Y$, which is blocked otherwise.
- **Example:** (right) A : family heart disease history; Y : smoking; L : heart disease; C : taking heart disease medication

Correlation and Causation



Correlation and Causation

In summary, there are three structural reasons why two variables may be correlated:

1. One causes the other[17]
2. They share common causes
3. The analysis is conditioned on their common effects[18]

[17] either directly or through mediating variables.
[18] or the consequences of the common effects.

Confounding

- When two variables share common causes, they are correlated even if they do not cause each other. This makes it harder for us to learn the causal effect one has on the other. We call this problem **confounding**. The common causes are called **confounders**.

- Self-selection bias is an important type of confounding: when patients choose hospitals based on their illness, illness is a common cause of both their treatment (hospital) and their outcome (recovery rate), and is therefore a confounder in the analysis of the causal effect of hospital treatment.

Confounding

- A basic strategy to deal with confounding is to condition on the common causes of treatment and outcome[19] (while avoiding controlling for any of their common effects).
  - Conditioning on common causes make two variables independent if they do not have direct causal effects on each other[20].
  - Therefore, any association between two variables after their common causes have been conditioned on should be due to causation[21].

---

[19] When we condition on a variable, we also say we control for the variable.

[20] i.e., these two variables should not be correlated *unless* there is causation.

[21] Another way to understand this strategy: after common causes are conditioned on, to the extent that individuals who receive different treatments are still different, the differences do not affect their outcomes. Hence, when we compare different treatment groups, we would be effectively comparing apples with apples.

The Back-Door Criterion

- More generally, if we can condition on a set of variables z that block all open noncausal paths[22] between treatment x and outcome y, then the causal effect of x on y is identified[23].
  - In this case, z is said to satisfy the **back-door criterion**[24].
  - Conditioning on z makes x **exogenous** to y[25].

[22] Noncausal paths between x and y are called **back-door paths**. These are the paths that, if left open, induce correlation between x and y that is not a result of x causing y.

[23] A causal effect is **identified** if it is possible to be estimated from observed data.

[24] We also need to make sure z does not contain variables that are the common effects of x and y.

[25] x is said to be **exogenous** to y if there is no open noncausal path between the two variables (and y does not cause x). Otherwise, x is **endogenous**.

The Back-Door Criterion

Given z that satisfies the back-door criterion, we have:

$$\mathbb{E}\left[y|\text{do}(x), z\right] = \mathbb{E}\left[y|x, z\right]$$

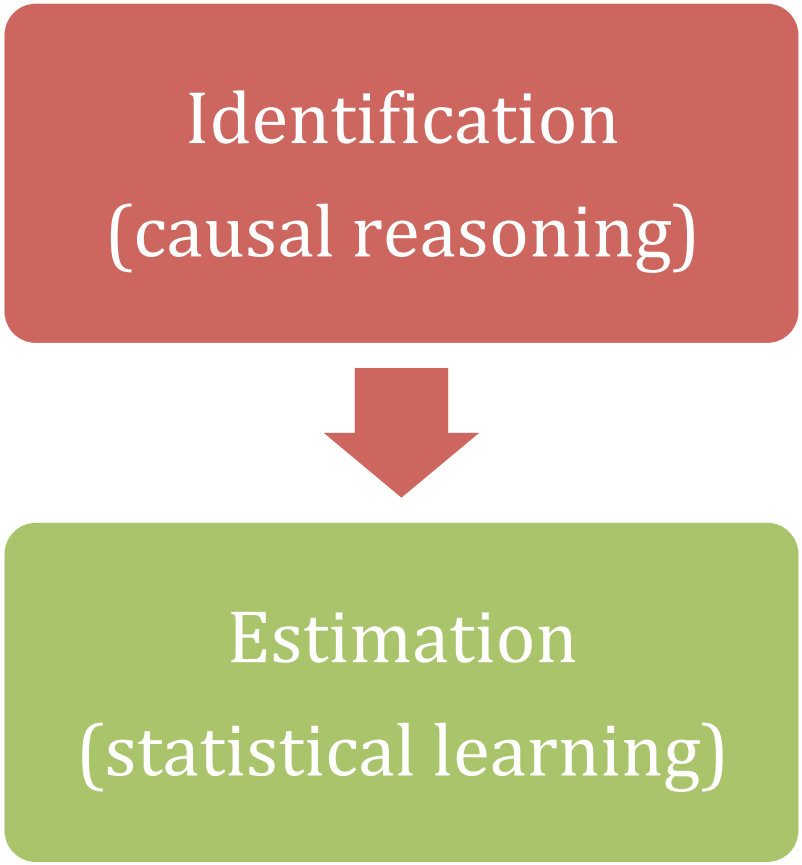
The average treatment effect[26]

$$\begin{aligned}
\text{ATE}\left(x\right) &\doteq \frac{d}{dx}\mathbb{E}\left[y|\text{do}(x)\right] \\
&= \frac{d}{dx}\mathbb{E}_z\left[\mathbb{E}\left[y|\text{do}(x), z\right]\right] \\
&= \frac{d}{dx}\mathbb{E}_z\left[\mathbb{E}\left[y|x, z\right]\right] = \mathbb{E}_z\left[\frac{\partial}{\partial x}\mathbb{E}\left[y|x, z\right]\right]
\end{aligned} \tag{1}$$

[26] When $x \in \{0, 1\}$ is binary,

$$\begin{aligned}
\text{ATE} &= \mathbb{E}\left[y|\text{do}(x = 1)\right] - \mathbb{E}\left[y|\text{do}(x = 0)\right] \\
&= \mathbb{E}_z\left[\mathbb{E}\left[y|x = 1, z\right] - \mathbb{E}\left[y|x = 0, z\right]\right]
\end{aligned}$$

Causal Effect Learning: Two Stages



Causal Effect Learning: Two Stages

- Once we have established identification based on causal reasoning, we can estimate the causal effect of interest using statistical models.

- Causal effect learning is therefore a two-stage process: in the first stage we determine what correlations in the data can tell us about causation (causal reasoning). In the second stage, we estimate the correlations from data (statistical learning).

Causal Effect Learning: Two Stages

- For example, suppose we have established that a set of observed variables z satisfies the back-door criterion, then according to (1), estimation of the ATE requires estimation of $\mathbb{E}[y|x,z]$.

- To estimate $\mathbb{E}[y|x,z]$ from data, we can use a variety of statistical models:
  - parametric or nonparametric
  - linear or non-linear

Causal Effect Learning: Two Stages

For simplicity, let

$$\mathbb{E}[y|x,z] = \beta_0 + \beta_1 x + \beta_2 z$$

Then

$$\widehat{\mathsf{ATE}} = \mathbb{E}_z\left[\frac{\partial}{\partial x}\widehat{\mathbb{E}}[y|x,z]\right] = \widehat{\beta}_1$$

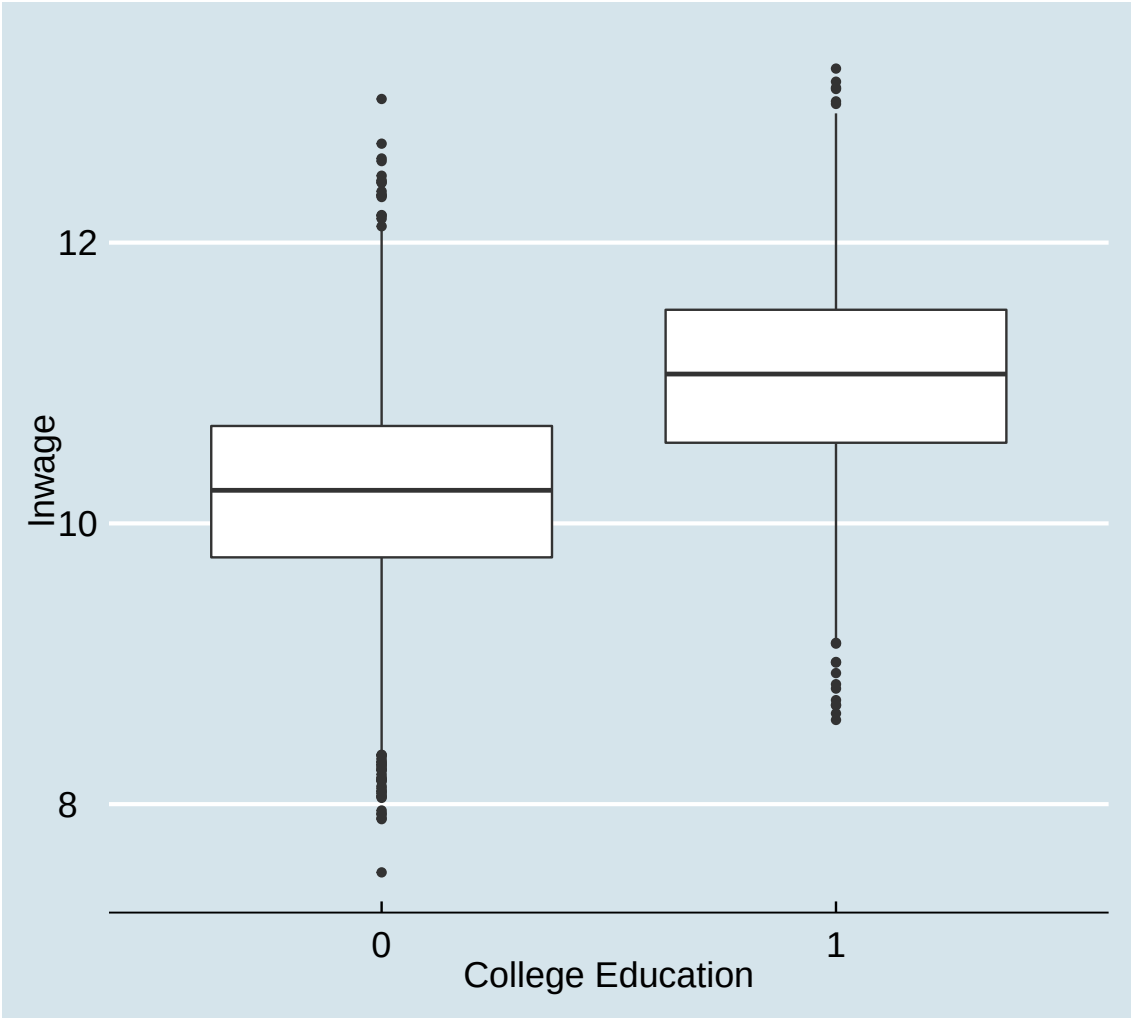
Returns to Education

Suppose we are interested in estimating the returns to college education. We have a sample of full-time employed individuals. Variables include demographic information, education level (non-college educated, college-educated), sector of employment, and current wage income.

```
data = read.csv('educ01.txt')
head(data)
```

```
##      lnwage sex age educ        sector
## 1 10.192260   1  32    0       Service
## 2  9.936083   0  36    0   Agriculture
## 3  9.248990   0  30    0   Agriculture
## 4 10.099410   1  36    0       Service
## 5  9.492408   1  31    0       Service
## 6 10.428530   0  32    0 Manufacturing
```

Returns to Education



Returns to Education

If there are no confounders to education and earnings – if education is exogenous to earnings, then

$$\mathbb{E}\left[\text{lnwage}|\, \text{do}\left(\text{Educ}\right)\right] = \mathbb{E}\left[\text{lnwage}|\, \text{Educ}\right] \tag{2}$$

Thus, we just need to estimate $\mathbb{E}\left[\text{lnwage}|\, \text{Educ}\right]$ from data[27].

Let[28]

$$\mathbb{E}\left[\text{lnwage}|\, \text{Educ}\right] = \beta_0 + \alpha \cdot \text{Educ} \tag{3}$$

, then

$$\begin{aligned}\text{ATE} &= \mathbb{E}\left[\text{lnwage}|\, \text{do}\left(\text{Educ} = 1\right)\right] - \mathbb{E}\left[\text{lnwage}|\, \text{do}\left(\text{Educ} = 0\right)\right] \\ &= \alpha\end{aligned}$$

[27] (2) is the identification (causal reasoning) step; (3) is the estimation (statistical modeling) step.

[28] When Educ is a continuous variable, there are many functional forms we can choose for modeling $\mathbb{E}[\text{lnwage}|\, \text{Educ}]$. When Educ is binary, (3) is both linear and nonparametric.

Returns to Education

$$\text{lnwage}_i = \beta_0 + \alpha \cdot \text{Educ}_i + e_i \tag{4}$$

```
require(AER)
fit.basic = lm(lnwage ~ educ, data=data)
coeftest(fit.basic)

##
## t test of coefficients:
##
##               Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 10.2205737  0.0084155 1214.490 < 2.2e-16 ***
## educ         0.8224472  0.0149943   54.851 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

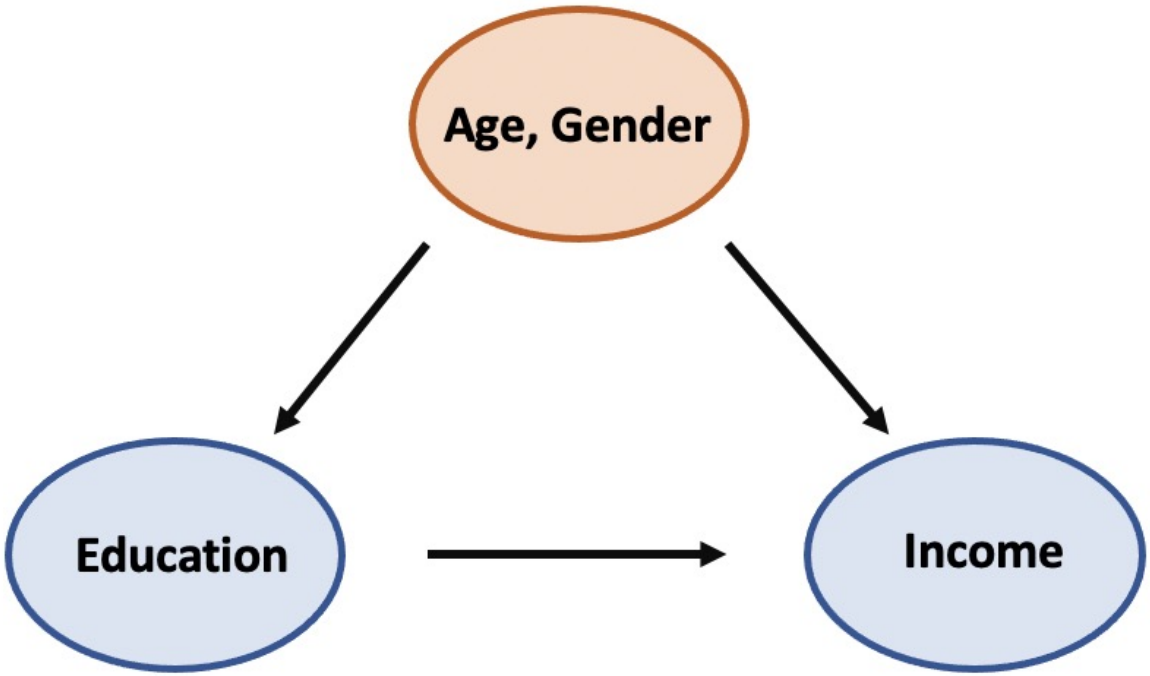
Adjusting for Observed Confounders

- The no confounding assumption amounts to assuming that individuals with and without college education are on average the same in all other aspects that could affect income.
- The ATE estimate ($\widehat{\alpha}$) we obtained from (4) is the same as the difference in mean log wage between college-educated and non-college-educated workers.
- But these individuals are different in some important measures that could affect income ...

Adjusting for Observed Confounders

- Gender
  - There can be gender differences in preference for higher education.
  - Men and women enjoy different labor market returns to education.
- Age:
  - Individuals born in different cohorts could have different preferences for higher education.
  - Age (work experience) is an important determinant of labor market returns.

Adjusting for Observed Confounders



Adjusting for Observed Confounders

If there are no other confounders, then education is exogenous to earnings conditional on age and sex. We have:

$$\mathbb{E}\left[\text{lnwage}|\,\text{do}\left(\text{Educ}\right),\text{sex},\text{age}\right]=\mathbb{E}\left[\text{lnwage}|\,\text{Educ},\text{sex},\text{age}\right]$$

Thus, we need to estimate $\mathbb{E}\left[\text{lnwage}|\,\text{Educ},\text{sex},\text{age}\right]$ from data.

Adjusting for Observed Confounders

Regression

Let[29]

$$\begin{aligned}
&\mathbb{E}\left[\text{lnwage}|\ \text{Educ}, \text{sex}, \text{age}\right] \\
&= \beta_0 + \alpha \cdot \text{Educ} + \beta_1 \cdot \text{sex} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{age}^2 \\
&= \alpha \cdot \text{Educ} + \beta \cdot X
\end{aligned}$$

, where $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$, $X = (\mathbf{1}, \text{sex}, \text{age}, \text{age}^2)$.

[29] Suppose sex is coded as a binary variable such that male $= 0$ and female $= 1$.

Adjusting for Observed Confounders

Regression

$$
\begin{aligned}
\text{ATE} &= \mathbb{E}\left[\text{lnwage}|\, \text{do}\left(\text{Educ}=1\right)\right] - \mathbb{E}\left[\text{lnwage}|\, \text{do}\left(\text{Educ}=0\right)\right] \\
&= \mathbb{E}_{\text{sex,age}}\left[\mathbb{E}\left[\text{lnwage}|\, \text{do}\left(\text{Educ}=1\right), \text{sex}, \text{age}\right]\right] \\
&\quad - \mathbb{E}_{\text{sex,age}}\left[\mathbb{E}\left[\text{lnwage}|\, \text{do}\left(\text{Educ}=0\right), \text{sex}, \text{age}\right]\right] \\
&= \mathbb{E}_{\text{sex,age}}\left(\mathbb{E}\left[\text{lnwage}|\, \text{Educ}=1, \text{sex}, \text{age}\right]\right. \\
&\qquad\qquad \left. -\mathbb{E}\left[\text{lnwage}|\, \text{Educ}=0, \text{sex}, \text{age}\right]\right) \\
&= \alpha
\end{aligned}
$$

Adjusting for Observed Confounders

Regression

$$\text{lnwage}_i = \alpha \cdot \text{Educ}_i + \beta \cdot X_i + e_i$$

```
fit.ols = lm(lnwage ~ educ + sex + poly(age,2,raw=T), data=data)
coeftest(fit.ols)

##
## t test of coefficients:
##
##                              Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)                8.13786695  0.22384811  36.3544 < 2.2e-16 ***
## educ                       0.81629233  0.01424411  57.3074 < 2.2e-16 ***
## sex                       -0.27702209  0.01321296 -20.9659 < 2.2e-16 ***
## poly(age, 2, raw = T)1     0.09377598  0.01234991   7.5933 3.399e-14 ***
## poly(age, 2, raw = T)2    -0.00084304  0.00016757  -5.0311 4.961e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adjusting for Observed Confounders

Matching

- In addition to regression, we could use **matching** to control for observed confounding when the treatment variable is discrete.

- If college education is assigned to each individual in a randomized trial, then $z = (\text{sex}, \text{age})$ would be independent of Educ:

$$p\left(z|\text{Educ} = 1\right) = p\left(z|\text{Educ} = 0\right) \tag{5}$$

, i.e., in a sample generated by randomized experiment, z is no longer a confounder and we can obtain the causal effect of college education by directly comparing the earnings of college-educated vs. non-collge-educated workers.

Adjusting for Observed Confounders

Matching

- The goal of matching is to construct a new sample (out of the observed sample) in which (5) holds true[30]. We can then treat the new sample – called the **matched sample** – *as if* it is generated by a randomized trial.

- The matched sample would allow us to compare apples with apples: college and non-college educated workers would have the same distribution of age and gender.

[30] Condition (5) is called **covariate balance**.

Adjusting for Observed Confounders

Matching

person	Gender	Age	Education	lnwage
1	female	young	1	11.3
2	male	young	1	11.7
3	female	old	1	11.4
4	female	old	0	10.3
5	male	old	0	10.7
6	male	young	0	10.5
7	female	young	0	10.6
8	male	old	0	10.5
9	male	young	0	10.6
10	male	old	0	11.3

Observed Sample

Adjusting for Observed Confounders

Matching

person	Gender	Age	Education	lnwage
1	female	young	1	11.3
2	male	young	1	11.7
3	female	old	1	11.4
4	female	old	0	10.3
6	male	young	0	10.5
7	female	young	0	10.6

Matched Sample

Adjusting for Observed Confounders

Matching

Matching estimator for the **average treatment effect on the treated (ATT)**[31,32]:

$$\begin{aligned} \text{ATT} &= \frac{1}{3}\left(11.3 + 11.7 + 11.4 - 10.3 - 10.5 - 10.6\right) \\ &= 1 \end{aligned}$$

[31] Notice that our matched sample has the same covariate distribution as the original treated population. Hence the effect we calculate on this sample is the ATT. To obtain the ATE in the observed sample, we would need to construct *two* matched samples. In the first one, we match non-college educated workers to college educated workers as we have done. This gives us the ATT. In the second one, we match college-educated workers to non-college educated workers. This would give us the **ATU** (**average treatment effect on the untreated**). Combining the two gives us the ATE.

[32] If the treatment effect is homogeneous, then ATE = ATT = ATU.

Adjusting for Observed Confounders

Matching

- When matching cannot be exact, we can match based on nearest neighbors.
- Matching could be performed *with* or *without* replacement.
  - When matching a group with fewer units to a group with more units, matching with replacement is necessary.

Adjusting for Observed Confounders

Matching

```
# Assess balance in the observed sample
require(tableone)
CreateTableOne(data, vars=c("sex","age"), strata="educ")
```

```
##                     Stratified by educ
##                      0             1             p      test
##   n                   6850          3150
##   sex (mean (SD))     0.53 (0.50)   0.44 (0.50)  <0.001
##   age (mean (SD))    35.25 (5.68)  34.57 (5.42)  <0.001
```

Adjusting for Observed Confounders

Matching

```
# Matching based on nearest neighbor
require(MatchIt)
m = matchit(educ ~ sex + age, data, distance="mahalanobis")
m$nn[c(-1,-3),]

##            Control Treated
## All           6850    3150
## Matched       3150    3150
## Unmatched     3700       0
## Discarded        0       0

# Create matched sample
m.data = match.data(m)
```

Adjusting for Observed Confounders

Matching

```
# Assess balance in the matched sample
CreateTableOne(m.data, vars=c("sex","age"), strata="educ")
```

```
##                     Stratified by educ
##                      0             1             p      test
##   n                   3150          3150
##   sex (mean (SD))    0.44 (0.50)   0.44 (0.50)   1.000
##   age (mean (SD))   34.57 (5.42)  34.57 (5.42)   1.000
```

Adjusting for Observed Confounders

Matching

```
# Matching estimator for ATT
fit.match = lm(lnwage ~ educ, data=m.data)
coeftest(fit.match)

##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 10.208691   0.012565 812.475 < 2.2e-16 ***
## educ         0.834330   0.017769  46.953 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Heterogeneous Treatment Effects

- We may be interested in how returns to college education differ by sector. Here sector of employment is a potential effect modifier.
- An **effect modifier** is a variable s such that given treatment x and outcome y[33],

$$\mathbb{E}\left[y|\text{do}\left(x\right),s\right] \neq \mathbb{E}\left[y|\text{do}\left(x\right)\right]$$

- Computing treatment effects separately for individuals with different values of s allows us to gauge how the effect of a treatment varies among the population:

$$\text{ATE}\left(x,s\right) = \frac{\partial}{\partial x}\mathbb{E}\left[y|\text{do}(x),s\right]$$

[33] Causally, many variables could be effect modifiers. An exogenous cause of the outcome variable whose effect interacts with that of the treatment is an effect modifier. A common cause to both treatment and outcome – a confounder — whose effect on the outcome interacts with that of the treatment is an effect modifier. A mediator that mediates the effect of the treatment on the outcome is also an effect modifer.

Heterogeneous Treatment Effects

```
require(dplyr)
require(broom)
data %>%
  group_by(sector) %>%
  do(tidy(lm(lnwage ~ educ + sex + poly(age,2,raw=T),
             data=.))[2,c(1,2,3)])

## # A tibble: 3 x 4
## # Groups:   sector [3]
##   sector        term  estimate std.error
##   <chr>         <chr>    <dbl>     <dbl>
## 1 Agriculture   educ     0.700    0.0271
## 2 Manufacturing educ     0.697    0.0150
## 3 Service       educ     0.759    0.0137
```

Unmeasured Confounding

- Age and Gender are not the only confounders in the relationship between education and earnings. One of the most important factor confounding this relationship is ability: individuals with higher abilities are more likely to attend and graduate from college, while they are also more likely to earn more regardless of educational attainment.

- Ability, however, is unobserved (if not ill-defined): it is an **unmeasured confounder**[34].

[34] In the econometrics and statistics literature, if all confounders are observed, we say there is **selection on observables** (or, there exists **no unmeasured confounding**). If some confounders are unobserved, we say there is **selection on unobservables** (or, there exists **unmeasured confounding**).

Unmeasured Confounding

- In the presence of unmeasured confounding, we can sometimes still find a set of observed variables that can block all noncausal paths between treatment and outcome and thus satisfy the back-door criterion[35].

- When this is not true, however, we need to find new ways to identify the causal effect of interest.

[35] See Appendix Figure 1. W is the confounder to X and Y, but we do not need to observe it: the causal effect of X on Y is identifiable by conditioning on C.

Instrumental Variables

- An **instrumental variable** (**IV**) is a variable that is (1) correlated with the treatment; (2) exogenous to the outcome; (3) affects the outcome *only* through its correlation with the treatment.

Instrumental Variables

Suppose we now observe parents' educational level (at least one college, no college) for each individual.

```
data = read.csv('educ02.txt')
head(data)

##      lnwage sex age educ        sector paeduc
## 1 10.192260   1  32    0       Service      0
## 2  9.936083   0  36    0   Agriculture      0
## 3  9.248990   0  30    0   Agriculture      0
## 4 10.099410   1  36    0       Service      0
## 5  9.492408   1  31    0       Service      0
## 6 10.428530   0  32    0 Manufacturing      0

cor(data$paeduc,data$educ)

## [1] 0.5211922
```

Instrumental Variables

- College educated parents can have positive impact on their children's college attainment, either through better home education or because they are more capable of affording college education.
- If we assume that more highly educated parents
  1. do not produce children that have higher unobserved abilities[36] or unobserved preferences[37] that affect education and earnings
  2. do not directly help their children obtain higher paying jobs

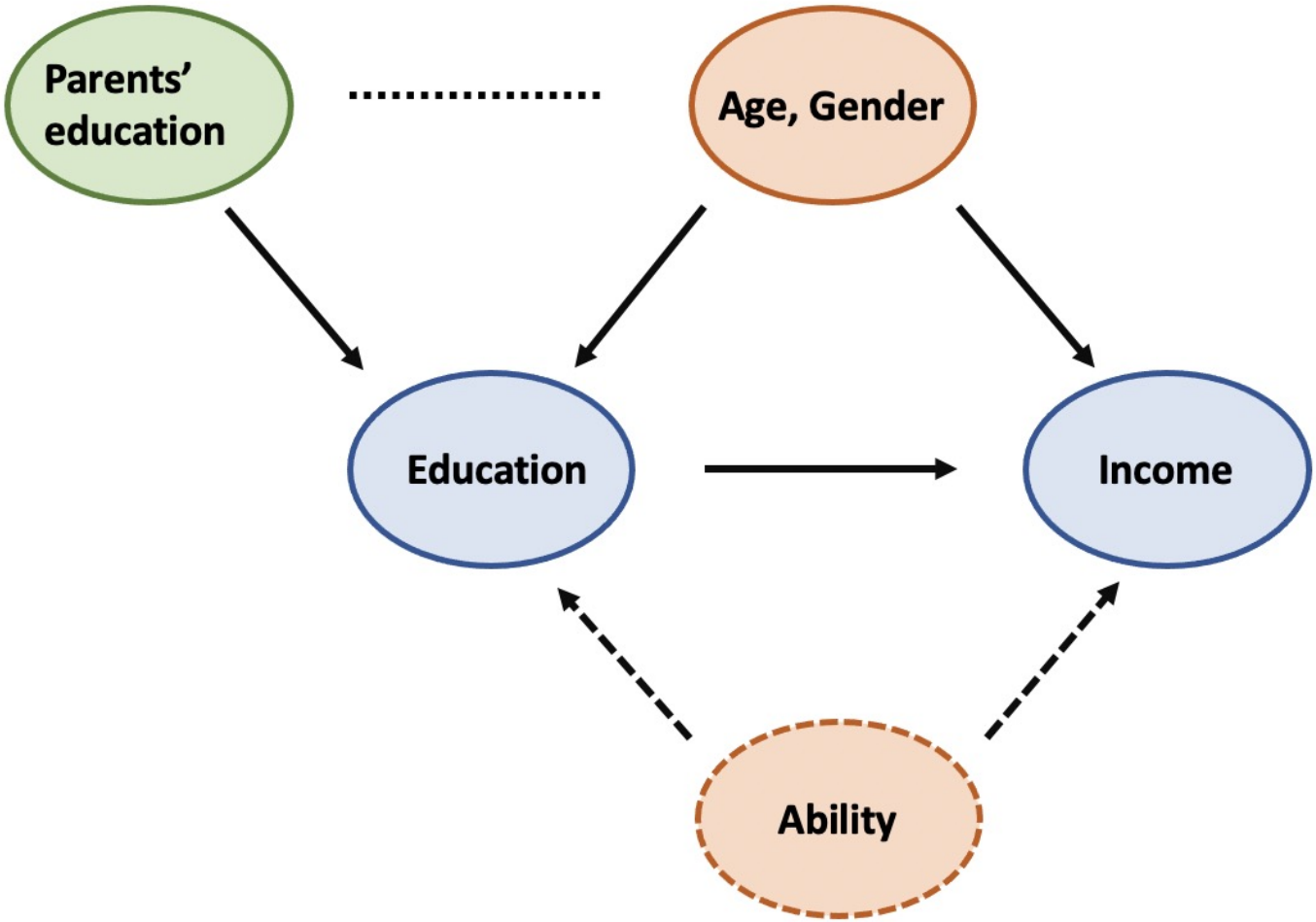
  Then the only way parents' education affects an individual's earnings is through its effect on her educational attainment[38]. Thus PaEduc can serve as an instrument for Educ.

[36] intelligence, social skills, good habits, etc.

[37] preference towards work, achievement, wealth, intellectual fulfillment, etc.

[38] In essence, we are assuming that more highly educated parents *only* help to increase their children's educational attainment, with no other discernible effects related to school and work.

Instrumental Variables



Instrumental Variables

If the effect of college education on income is *linearly separable* from the effect of other factors U (age, gender, ability, etc.):

$$\text{lnwage} = \alpha \cdot \text{Educ} + U$$

Then

$$\begin{aligned} Cov\left(\text{lnwage}, \text{PaEduc}\right) &= Cov\left(\alpha \cdot \text{Educ} + U, \text{PaEduc}\right) \\ &= \alpha \cdot Cov\left(\text{Educ}, \text{PaEduc}\right) \end{aligned}$$

$\Rightarrow$

$$\alpha = \frac{Cov\left(\text{lnwage}, \text{PaEduc}\right)}{Cov\left(\text{Educ}, \text{PaEduc}\right)}$$

Instrumental Variables

```
# if we believe paeduc may be correlated with age or gender
# for example, parents of older individuals tend to have lower
# education due to cohort effects, then we can use paeduc as instrument
# for educ conditional on these variables.

fit.iv = ivreg(lnwage ~ educ + sex + poly(age,2,raw=T) |
                paeduc + sex + poly(age,2,raw=T), data=data)
coeftest(fit.iv)

##
## t test of coefficients:
##
##                              Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)                 8.26375807  0.22851155  36.1634 < 2.2e-16 ***
## educ                        0.53184251  0.02782403  19.1145 < 2.2e-16 ***
## sex                        -0.30148429  0.01362766 -22.1230 < 2.2e-16 ***
## poly(age, 2, raw = T)1      0.09372078  0.01259387   7.4418 1.075e-13 ***
## poly(age, 2, raw = T)2 -0.00086055  0.00017088  -5.0359 4.838e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Instrumental Variables

- In this IV strategy, the source of exogenous variation comes from PaEduc not Educ.
- To obtain the treatment effect of Educ on lnwage, we "net out" the effect of PaEduc on Educ *from* the effect of PaEduc on lnwage.
- But what if some individuals are *always* going to college, while some are *never* going to college, regardless of their parents' educational level?[39] The treatment effect of college education for *these* people would be not identified.

[39] For example, smart children living in areas with good public schools may have a high probability of going to college regardless of their parents' educational level.

Instrumental Variables

- α can be understood as the effect of Educ on lnwage for those whose Educ changes *in response to* changes in PaEduc.

- When a treatment effect is heterogeneous among the population, the IV strategy identifies a **local average treatment effect** (**LATE**) – the average treatment effect among those whose treatment status changes in response to or in association with the instrument[40].

[40] If the treatment effect is homogeneous, then LATE = ATE.

Fixed Effects

Suppose now we know that our sample of individuals are collected from different cities.

```
data = read.csv('educ03.txt')
head(data)
```

```
##      lnwage sex age educ        sector paeduc city
## 1 10.192260   1  32    0       Service      0    1
## 2  9.936083   0  36    0   Agriculture      0    1
## 3  9.248990   0  30    0   Agriculture      0    1
## 4 10.099410   1  36    0       Service      0    1
## 5  9.492408   1  31    0       Service      0    1
## 6 10.428530   0  32    0 Manufacturing      0    1
```

```
# number of cities
length(unique(data$city))
```

```
## [1] 50
```

Fixed Effects

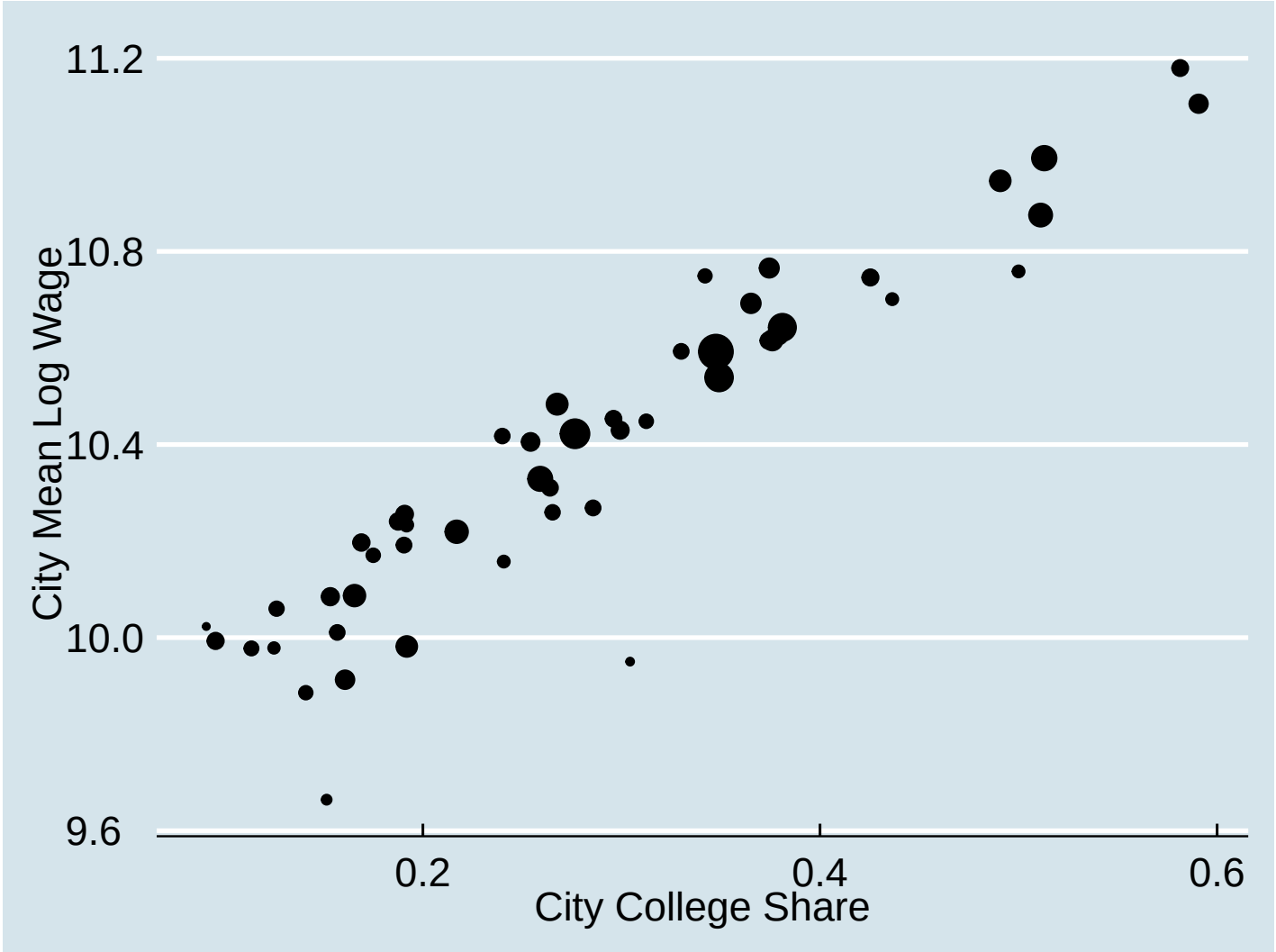
What might be problems?

- Higher income cities have better-paying jobs and better schools[41] – individuals who attend schools in these cities are more likely to go to college, and are more likely to earn higher wages upon graduation.
- Higher income cities may through migration attract individuals who have higher ability and better education[42].

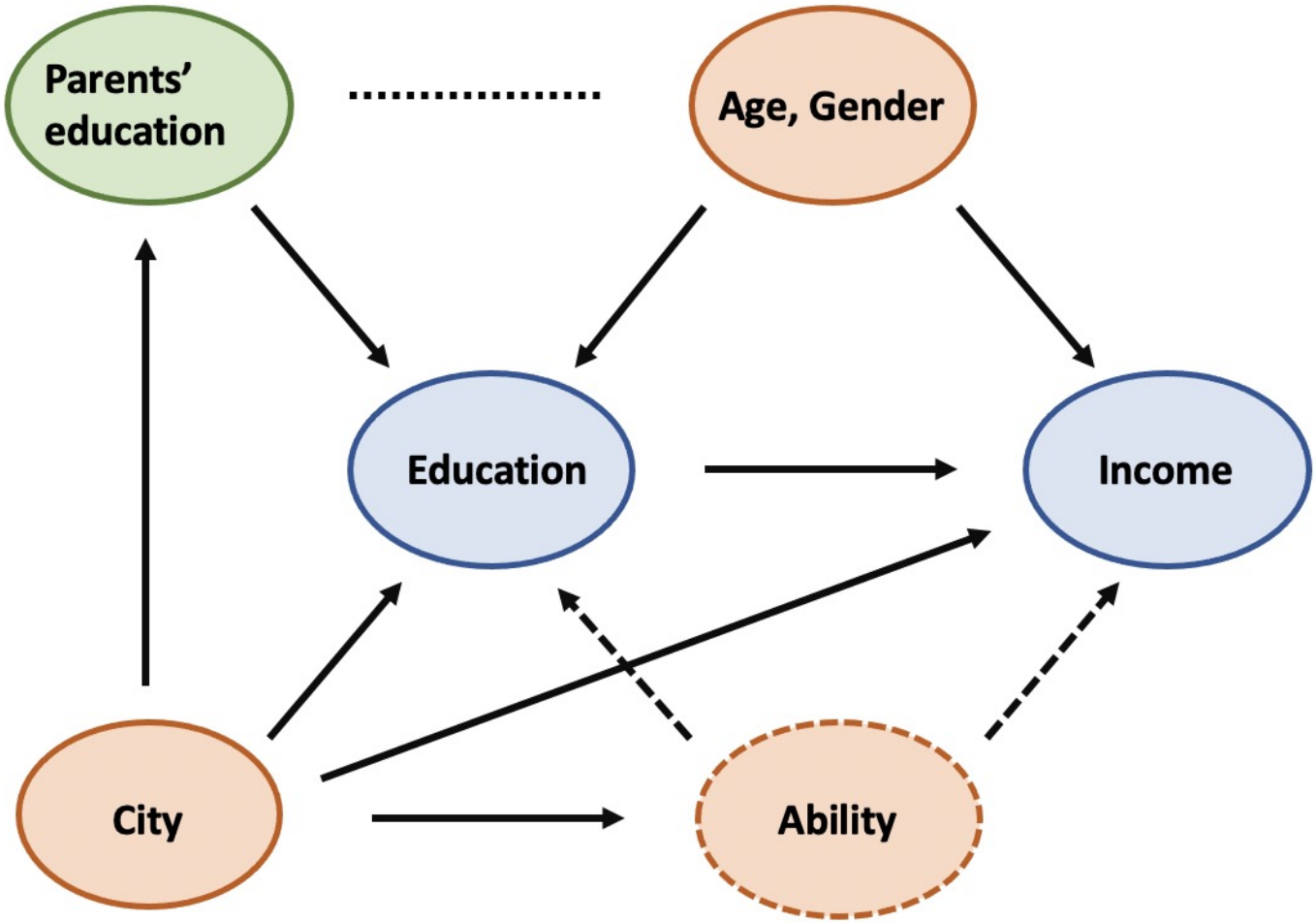
[41] elementary, middle, and high schools

[42] For the same reasons, parents of individuals in higher income cities also tend to have higher educational attainment. To the extent that parents' education affects children's education, their children will be more likely to go to college and – upon graduation, if working in the same high-income cities – earn higher wages.

Fixed Effects



Fixed Effects



Fixed Effects

While ability, school quality, and productivity are all unobserved, if we assume they are the same for each individual within a city – if we assume these variables mainly vary at the city level – then we can control for them by treating city itself as a confounder:

$$\mathbb{E}\left[\text{lnwage}\middle|\,\text{do}\left(\text{Educ}\right),\text{sex},\text{age},\text{city}\right]=\mathbb{E}\left[\text{lnwage}\middle|\,\text{Educ},\text{sex},\text{age},\text{city}\right]$$

Statistically, let i denote individual and m denote city. Let

$$\text{lnwage}_{i,m}=\tau_m+\alpha\cdot\text{Educ}_{i,m}+\beta\cdot X_{i,m}+e_{i,m}$$

Then τ_m are called city **fixed effects** and α is our desired ATE.

Fixed Effects

```
require(lfe)
fit.fe = felm(lnwage ~ educ + sex + poly(age,2,raw=T) | city, data)
coeftest(fit.fe)

##
## t test of coefficients:
##
##                             Estimate  Std. Error  t value  Pr(>|t|)
## educ                     0.69942765  0.01390108  50.3146 < 2.2e-16 ***
## sex                     -0.28637255  0.01251160 -22.8886 < 2.2e-16 ***
## poly(age, 2, raw = T)1   0.09507713  0.01168000   8.1402 4.420e-16 ***
## poly(age, 2, raw = T)2 -0.00086538  0.00015848  -5.4604 4.865e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fixed Effects

```
# Since ability varies mainly at the individual rather than city level
# (even though high-income cities may on average attract higher ability
# individuals), a city fixed effect may not eliminate ability
# confounding. In this case, we can combine fixed effect modeling
# with instrumental variables: Conditional on city, parents' education
# remains a valid IV.

fit.feiv = felm(lnwage ~ sex + poly(age,2,raw=T) |
                city | (educ ~ paeduc), data)
coeftest(fit.feiv)

##
## t test of coefficients:
##
##                          Estimate  Std. Error  t value  Pr(>|t|)
## sex                   -0.30502075  0.01281397 -23.8038 < 2.2e-16 ***
## poly(age, 2, raw = T)1  0.09487645  0.01182513   8.0233 1.146e-15 ***
## poly(age, 2, raw = T)2 -0.00087629  0.00016046  -5.4613 4.842e-08 ***
## `educ(fit)`             0.48021022  0.02674655  17.9541 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Quasi Experiments

- In non-experimental settings, circumstances sometimes produce what appears to be randomization.
- Because of man-made rules or external events, the treatment of some individual occurs *as if* it is random random.
- Such "as if" randomness produces a **quasi-experiment** or **natural experiment**. Causal inference strategies based on exploiting such "as if" randomness are called **quasi-experimental designs**.

Regression Discontinuity Design

Suppose now we have information on each individual's college entrance exam score (above or below admission cutoff).

```
data = read.csv('educ04.txt')
head(data)
```

```
##      lnwage sex age educ        sector paeduc city test
## 1 10.192260   1  32    0       Service      0    1  -46
## 2  9.936083   0  36    0   Agriculture      0    1  -35
## 3  9.248990   0  30    0   Agriculture      0    1  -49
## 4 10.099410   1  36    0       Service      0    1   -7
## 5  9.492408   1  31    0       Service      0    1  -38
## 6 10.428530   0  32    0 Manufacturing      0    1  -35
```

Regression Discontinuity Design

- Admission cutoff provides a natural experiment on college attendance.

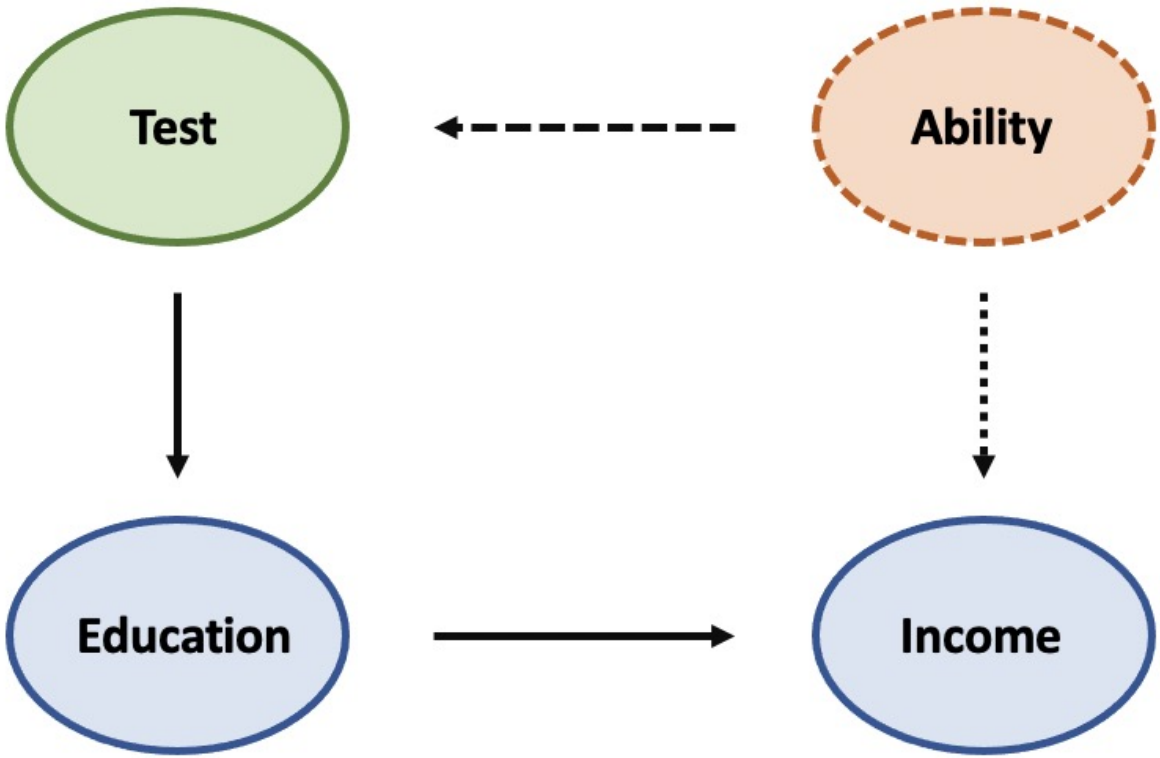
- Students just above or below an admission cutoff are likely to be very similar on observable and unobservable characteristics. Due to chance variation (perhaps due to how they feel on the exam day), those who are above the cutoff have the opportunity to go to college, while those are below do not. For these students, college attendance is *as if* random. Comparing them produces an estimate of the causal effect of college education.

Regression Discontinuity Design

At the heart of the **regression discontinuity design** (**RDD**) are discrete treatment status being determined by an underlying continuous **running variable**:

$$\text{Educ} = \begin{cases} 1 & \text{if test} \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

Regression Discontinuity Design



Regression Discontinuity Design

The causal diagram implies that

$$\mathbb{E}\left[\text{lnwage}|\, \text{do}\left(\text{Educ}\right), \text{test}\right] = \mathbb{E}\left[\text{lnwage}|\, \text{Educ}, \text{test}\right]$$

But there is no way to compare $\mathbb{E}\left[\text{lnwage}|\, \text{Educ} = 1, \text{test} = c\right]$ with $\mathbb{E}\left[\text{lnwage}|\, \text{Educ} = 0, \text{test} = c\right]$[43], since at any $c \neq 0$ we either only have Educ $= 1$ or Educ$=0$.

- In this case we have a lack of **overlap**.

[43] unless we rely on extrapolation based on functional form assumptions.

Regression Discontinuity Design

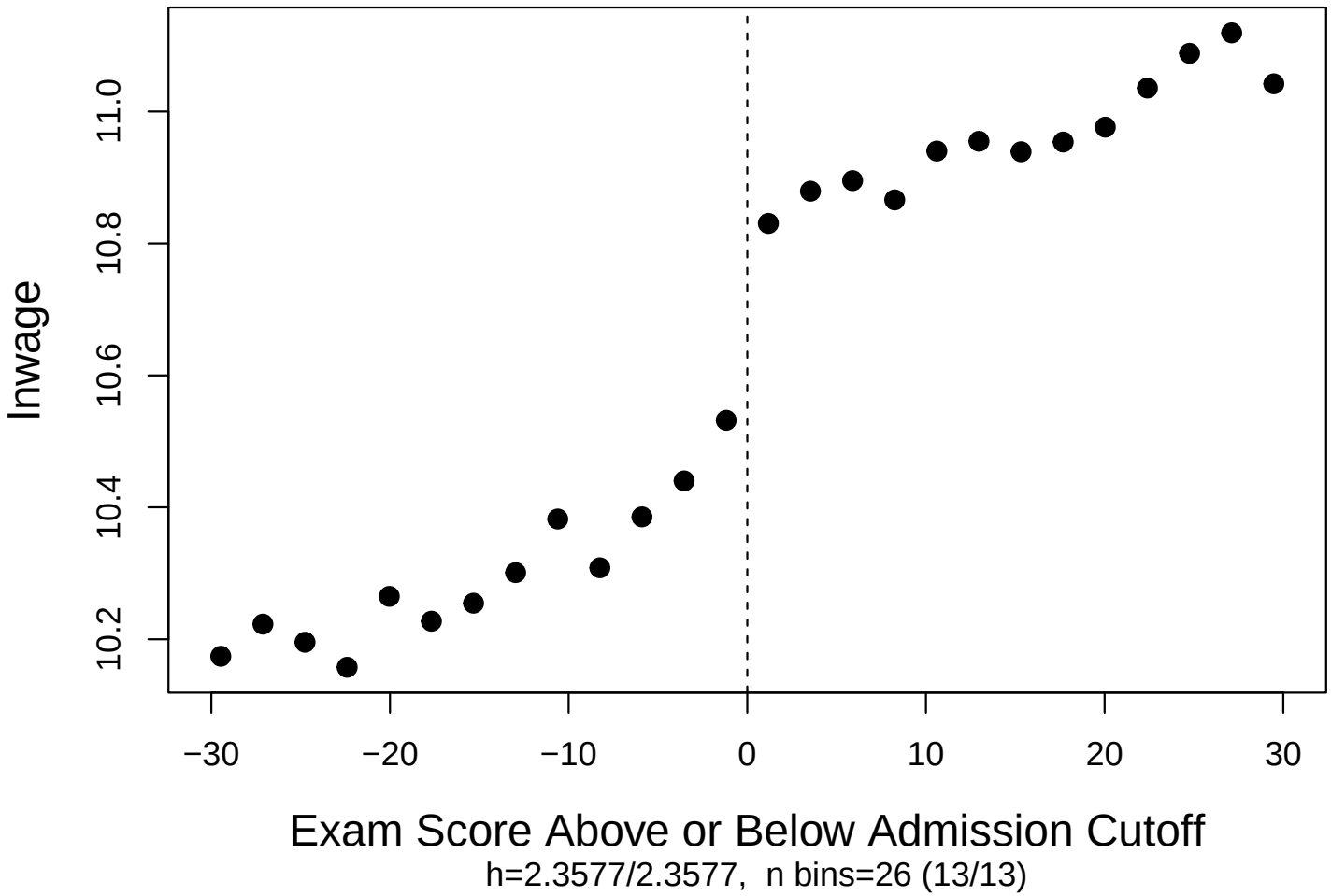
Only in a small neighborhood around test $= 0$ do we observe both college and non-college educated individuals. Hence we can compare $\mathbb{E}[\text{lnwage}|\,\text{Educ} = 1, \text{test}]$ with $\mathbb{E}[\text{lnwage}|\,\text{Educ} = 0, \text{test}]$ for individuals in this small neighborhood.

The result is an estimate of a LATE – average treatment effect for those individuals at the cutoff. Formally,

$$\begin{aligned}\text{LATE} &= \lim_{\text{test}\to 0_+} \mathbb{E}[\text{lnwage}|\,\text{Educ} = 1, \text{test}] \\ &\quad - \lim_{\text{test}\to 0} \mathbb{E}[\text{lnwage}|\,\text{Educ} = 0, \text{test}] \\ &\overset{[1]}{=} \lim_{\text{test}\to 0_+} \mathbb{E}[\text{lnwage}|\,\text{test}] - \lim_{\text{test}\to 0} \mathbb{E}[\text{lnwage}|\,\text{test}]\end{aligned}$$

, where [1] follows since $\text{Educ} = \mathcal{I}(\text{test} \geq 0)$.

Regression Discontinuity Design



Regression Discontinuity Design

```
require(rddtools)
rdd.data = rdd_data(y=data$lnwage, x=data$test, cutpoint=0)
fit.rdd = rdd_reg_lm(rdd.data)
coeftest(fit.rdd)

##
## t test of coefficients:
##
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 1.0462e+01 1.6606e-02 629.9938   <2e-16 ***
## D           3.2915e-01 2.6519e-02  12.4120   <2e-16 ***
## x           1.0795e-02 5.9674e-04  18.0895   <2e-16 ***
## x_right     9.6842e-04 1.3256e-03   0.7305   0.4651
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression Discontinuity Design

In practice, college attendance is not completely determined by exam scores in our data.

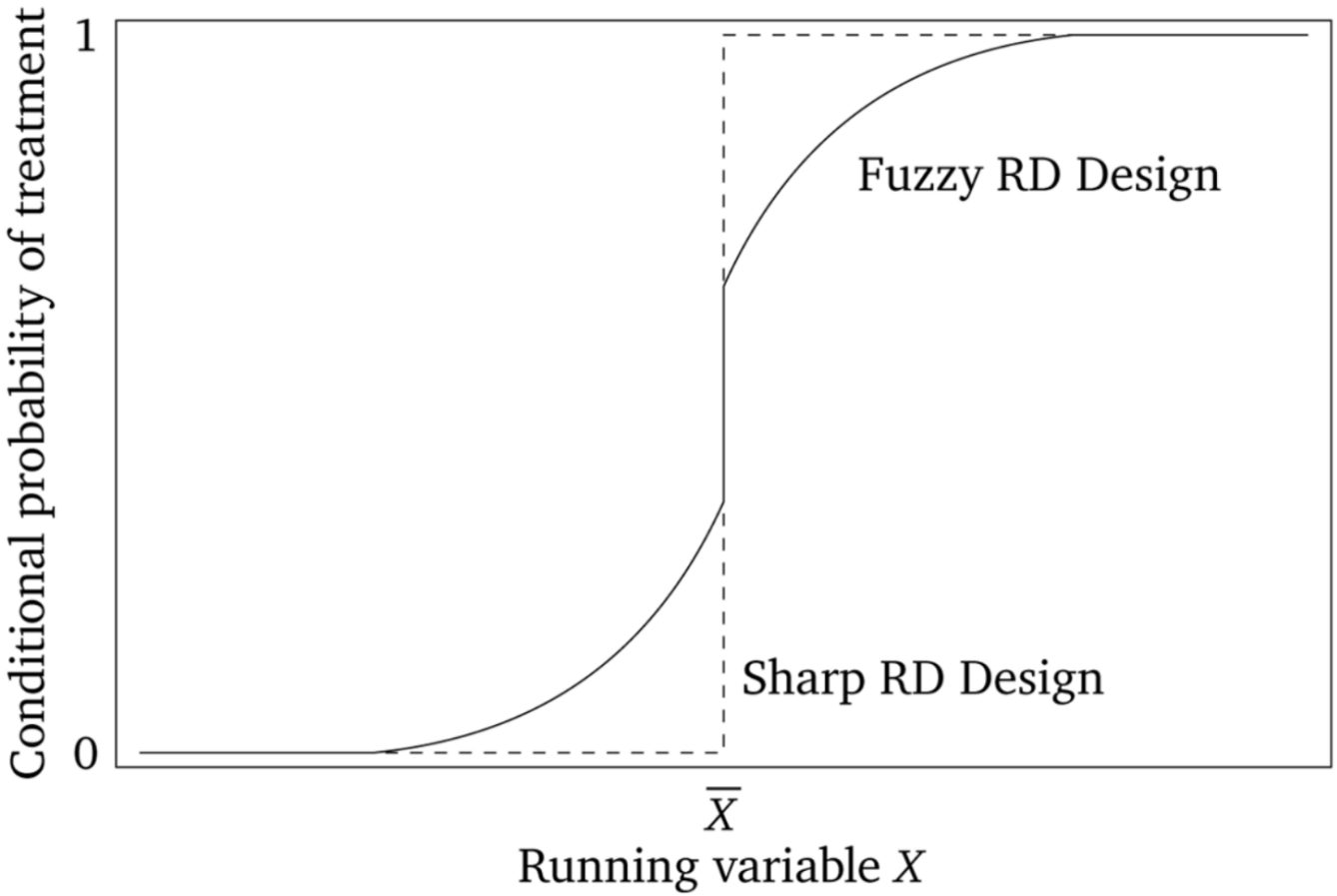
```
aggregate(educ ~ (test>0),data,mean)
```

```
##   test > 0        educ
## 1    FALSE 0.009368761
## 2     TRUE 0.885353970
```

Instead, an admission cutoff is associated with a discontinuous jump in the probability of college attendance. This is called a **fuzzy RDD**[44].

[44] In contrast, an RDD in which treatment status is a deterministic function of the running variable is called a **sharp RDD**.

Regression Discontinuity Design



Regression Discontinuity Design

- Because college education is not entirely determined by exam scores, conditional on exam score, the college attendance decision can still be an endogenous one.

- Let $z = \mathcal{I}(\text{test} \geq 0)$. If we assume that in a small neighborhood around test $= 0$, z is almost exogenous, then we can use it as an instrument for Educ for individuals in that neighborhood to identify their local treatment effect.

- In general, when a quasi-experiment partially determines the treatment status, its "as if" randomness can be used as an instrument for identifying the causal effect of interest.

Regression Discontinuity Design

```
fzrdd.data = rdd_data(y=data$lnwage, x=data$test, cutpoint=0, z=data$educ)
fit.fzrdd = rdd_reg_lm(fzrdd.data)
coeftest(fit.fzrdd)

##
## t test of coefficients:
##
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept) 10.46177322  0.01638770 638.392   <2e-16 ***
## D            0.41754974  0.03319830  12.577   <2e-16 ***
## x            0.01079469  0.00058889  18.331   <2e-16 ***
## x_right     -0.00135111  0.00138153  -0.978   0.3281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression Discontinuity Design

```
# we can also control for other observed confounders and exogenous
# variables (X). One way to do this is to first regress lnwage on X
# and obtain its residuals (y). Doing so allows us to "parcel out"
# the effect of X on lnwage. We can then estimate the causal effect of
# Educ on y based on a (fuzzy) RDD.

fit = lm(lnwage ~ factor(city) + sex + poly(age,2,raw=T), data=data)
y = fit$residuals
fzrddX.data = rdd_data(y=y, x=data$test, cutpoint=0, z=data$educ)
fit.fzrddX = rdd_reg_lm(fzrddX.data)
coeftest(fit.fzrddX)

##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -0.07882535  0.01486951 -5.3011 1.175e-07 ***
## D            0.46886185  0.03012275 15.5650 < 2.2e-16 ***
## x            0.00590748  0.00053433 11.0558 < 2.2e-16 ***
## x_right     -0.00200588  0.00125354 -1.6002    0.1096
## ---
```

Difference-in-Differences

- Suppose now we observe a repeated cross section of individuals in $M = 50$ cities for $T = 10$ years.
- In addition, we know that at year 5, several cities started a college tuition subsidy program intended to help students afford college. The presence of such a program is coded as policy $\in \{0, 1\}$.

Difference-in-Differences

```
data = read.csv('educ05.txt')
head(data)
```

```
##   t city    lnwage sex age educ        sector paeduc policy
## 1 1    1 10.192260   1  32    0       Service      0      0
## 2 1    1  9.936083   0  36    0   Agriculture      0      0
## 3 1    1  9.248990   0  30    0   Agriculture      0      0
## 4 1    1 10.099410   1  36    0       Service      0      0
## 5 1    1  9.492408   1  31    0       Service      0      0
## 6 1    1 10.428530   0  32    0 Manufacturing      0      0
```

```
# number of cities that implemented the program
length(unique(data$city[data$policy==1]))
```

```
## [1] 9
```

Difference-in-Differences

Let's first discuss whether such a program is effective in promoting college attendance.

- If we compare a city that implemented the program before and after year 5, the change in its college attainment rate may be due to factors other than the subsidy program.
  - For example, the educational level of a city's population may be naturally rising even in the absence of the subsidy program.
- If we compare cities that implemented the program with those that did not after year 5, our comparison will be biased if the two groups of cities are different (selection bias).
  - For example, the cities that implemented the program could have lower human capital.

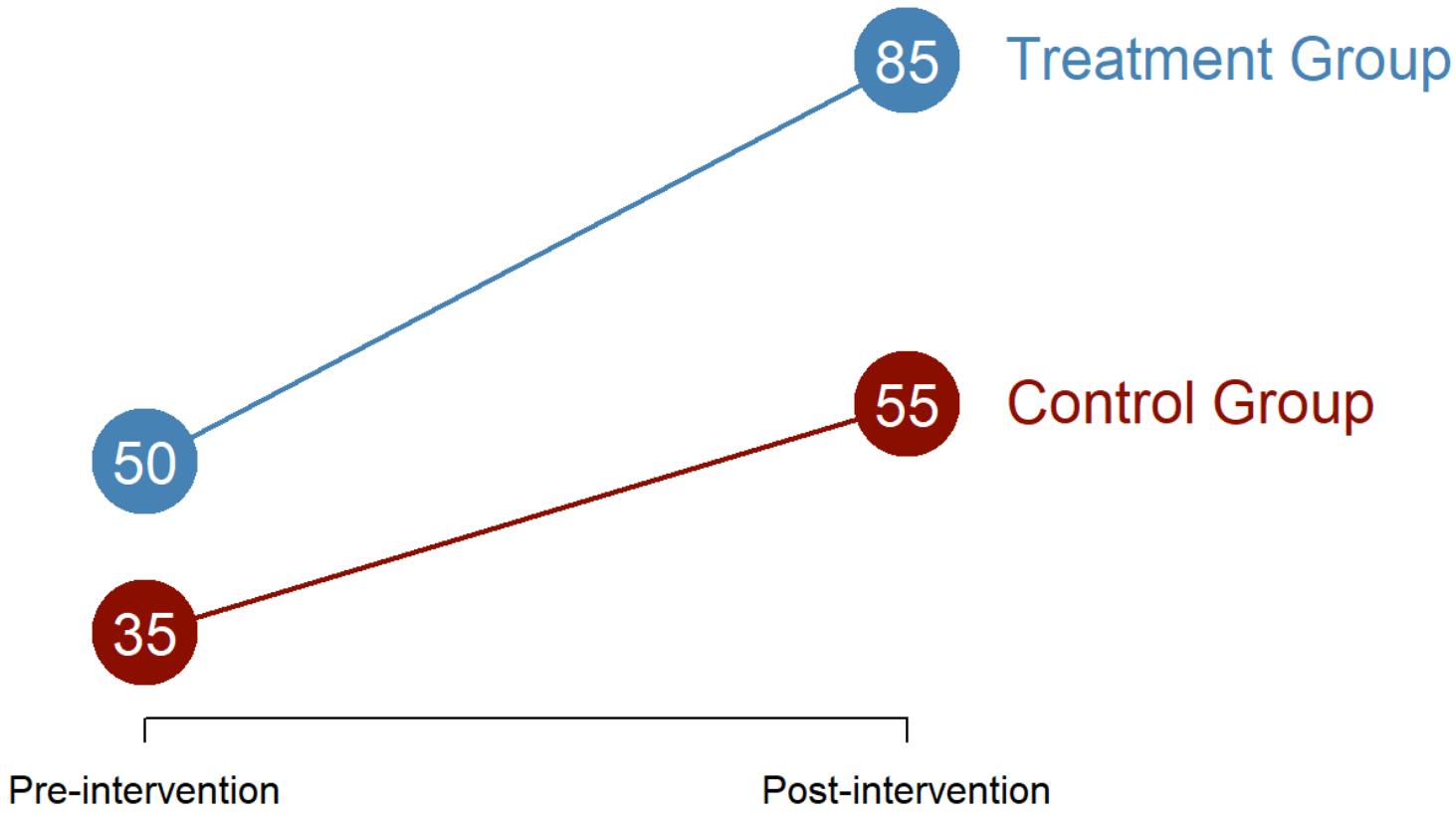
Difference-in-Differences

- If we assume, however, that *in the absence of the program*, the cities that implemented the program (**treatment group**) and the cities that did not implement the program (**control group**) are different in the level of their college attainment rate but similar in trend over time[45,46], then we can identify the causal effect of the subsidy program by comparing the change in college attainment rate of the treatment group vs. the control group before and after program implementation.

[45] due to being subject to the same factors that affect a city's educational level over time, say, national increase in the demand for high-skill workers.

[46] In essence, we are assuming that, absent treatment, the *difference* between the treatment and the group group is time-invariant. Thus, any *difference* in their *difference* must be due to the treatment effect.

Difference-in-Differences



The Difference-in-Differences Design

Difference-in-Differences

Let $\overline{\text{Educ}}$ denote the college attainment rate of a city. Let pre and post denote pre-program (year≤ 5) and post-program (year> 5).

Given the **parallel trend assumption**, the **difference-in-differences** (**DID**) estimate of the causal effect of the program is[47]

$$\begin{aligned}\alpha = &\left(\mathbb{E}\left[\overline{\text{Educ}}\middle|\,\text{treated}, \text{post}\right] - \mathbb{E}\left[\overline{\text{Educ}}\middle|\,\text{treated}, \text{pre}\right]\right) \\ &- \left(\mathbb{E}\left[\overline{\text{Educ}}\middle|\,\text{control}, \text{post}\right] - \mathbb{E}\left[\overline{\text{Educ}}\middle|\,\text{control}, \text{pre}\right]\right)\end{aligned} \tag{6}$$

[47] Technically, α is an ATT because the parallel trend assumption assumes what the treated cities would be like in the absence of the program, not what the control cities would be like given the program.

Difference-in-Differences

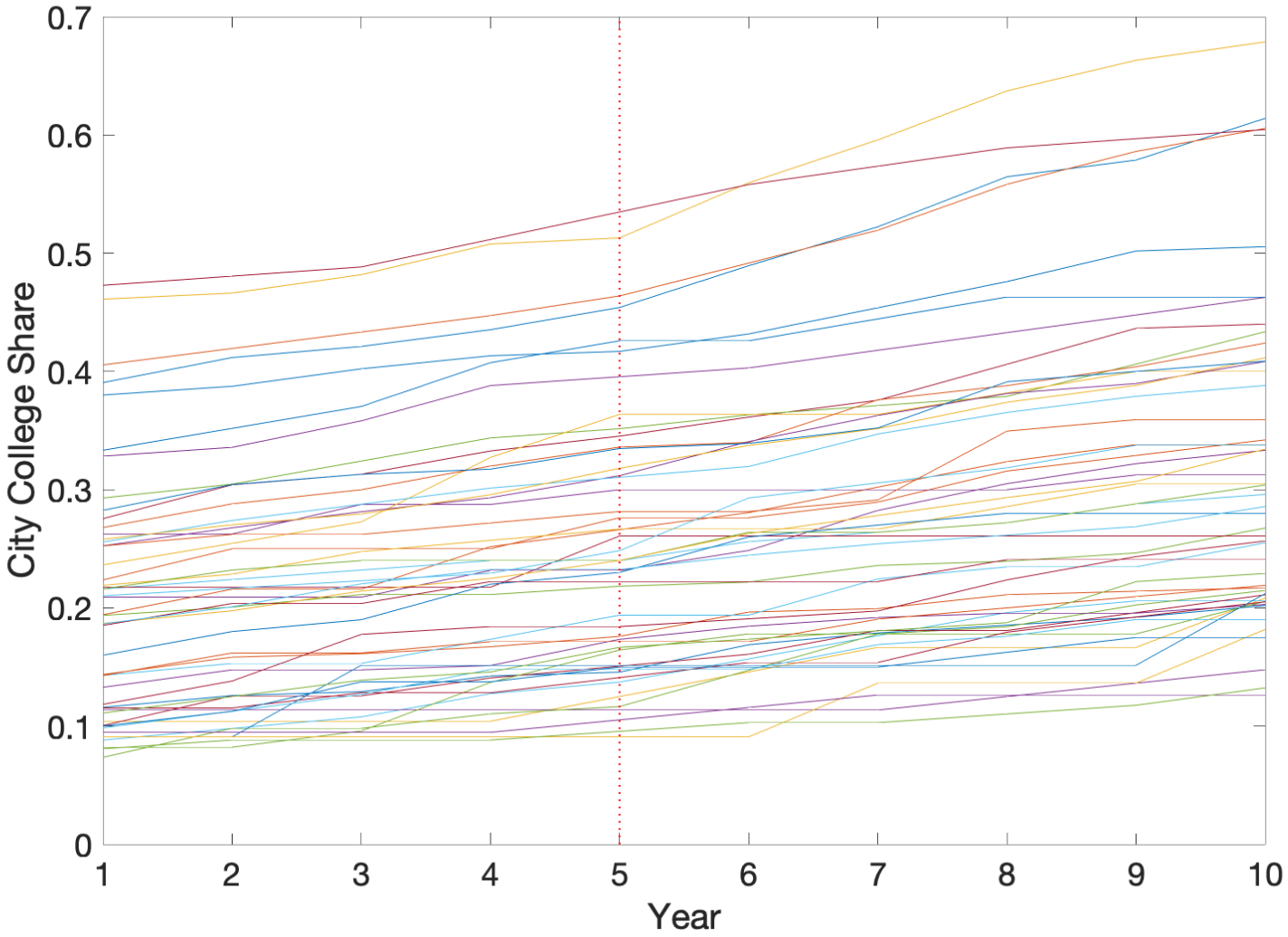
Equivalently, we can obtain α by estimating the following model[48]:

$$\overline{\text{Educ}}_{m,t} = \tau_m + \lambda_t + \alpha \cdot \text{policy}_{m,t} + e_{m,t} \tag{7}$$

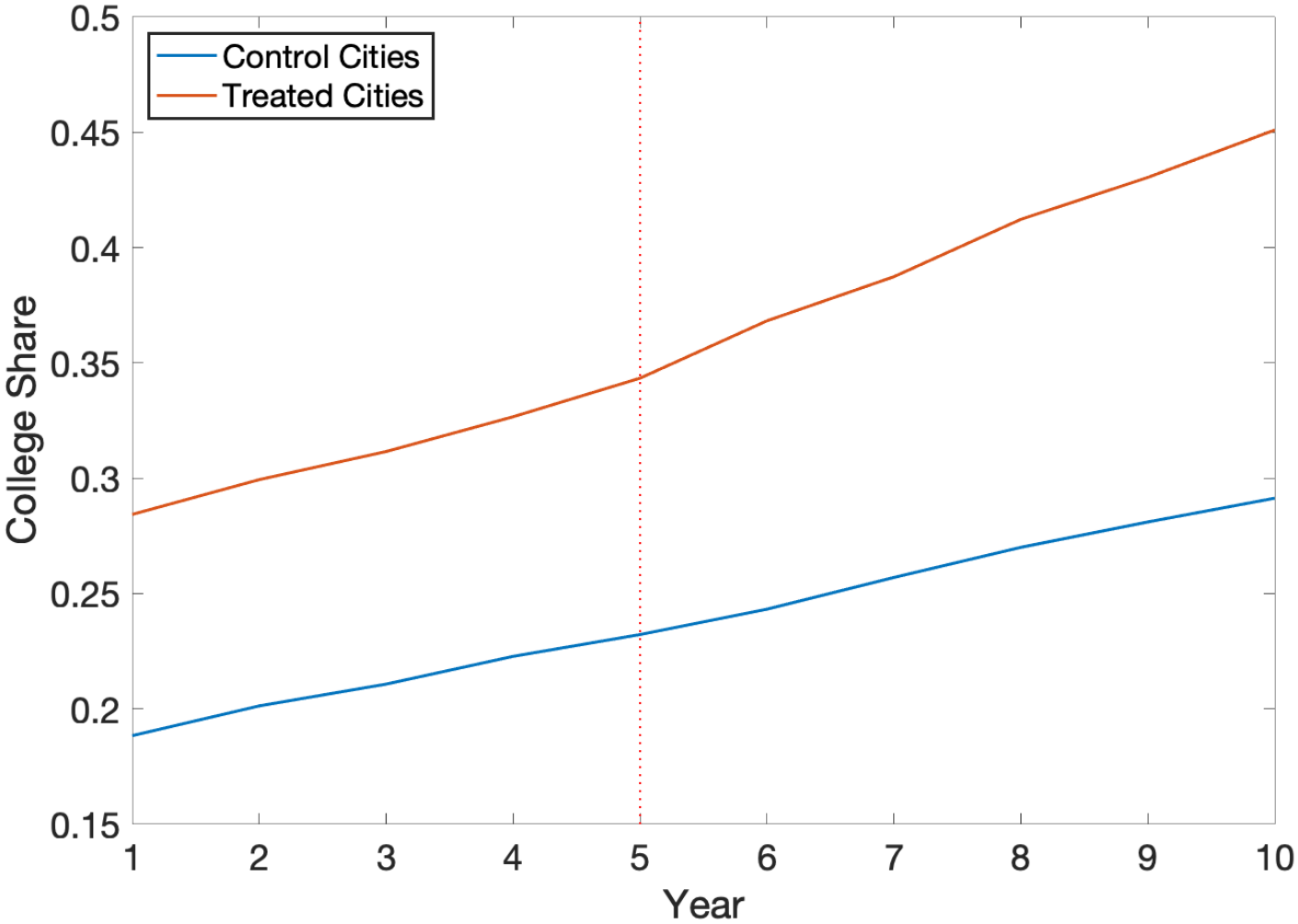
, where τ_m and λ_t are city and year fixed effects, and $\text{policy}_{m,t} = \mathcal{I}\,(m \in \text{treated}\ \&\ t \in \text{post})$.

[48] The fixed effect model satisfies the parallel trend assumption because differences among units (τ_m) are time-invariant, while all units evolve according to the same time trend (λ_t) in the absence of treatment.

Difference-in-Differences



Difference-in-Differences



Difference-in-Differences

```r
clist = unique(data$city[data$policy==1]) # treated cities
data$g = (data$city %in% clist) # g=0 if control; g=1 if treated
data$p = (data$t > 5) # p=0 if pre; p=1 if post

D = data %>% group_by(p,g) %>% summarise(educ = mean(educ))

ATT = (D$educ[D$g==1 & D$p==1] - D$educ[D$g==1 & D$p==0]) -
      (D$educ[D$g==0 & D$p==1] - D$educ[D$g==0 & D$p==0])
ATT
```

```
## [1] 0.03935337
```

Difference-in-Differences

```
coeftest(felm(educ ~ policy | t + city, data))

##
## t test of coefficients:
##
##          Estimate Std. Error t value  Pr(>|t|)
## policy 0.0393534  0.0056219  7.0001 2.574e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

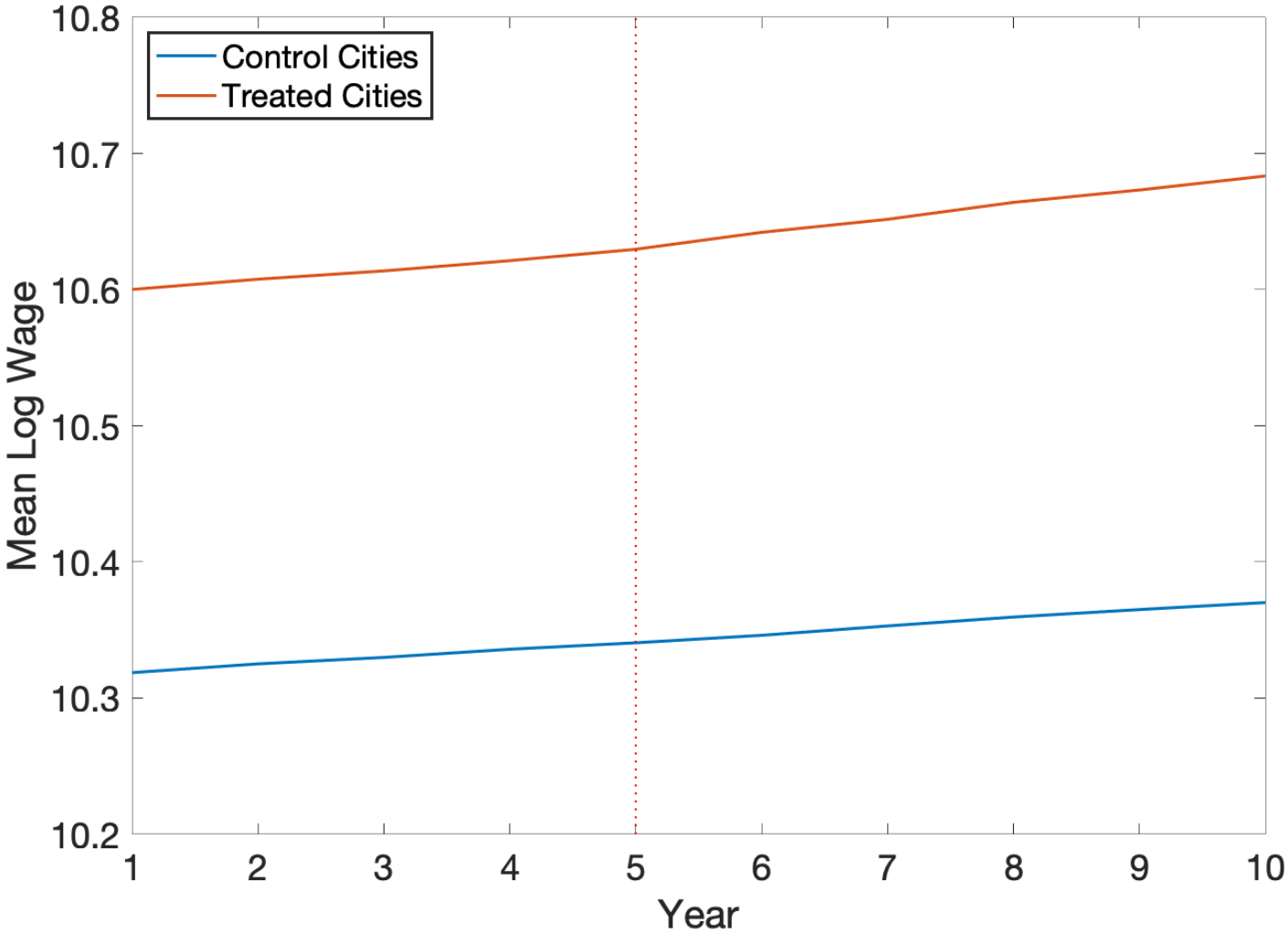
Difference-in-Differences

- Now let's come back to the problem of evaluating the causal effect of college education on income.

- If, conditional on city and year, the implementation of the subsidy program is exogenous to individual earnings – the program is a quasi-experiment – then, since we know it promotes college attainment, it can serve as an instrument for Educ in estimating its effect on lnwage.

- This requires another parallel trend assumption: absent the program, cities in the treatment group would have the same trend in *income* as those in the control group.

Difference-in-Differences



Difference-in-Differences

```
# here we also include individual covariates (observed confounders)
fit.did = felm(lnwage ~ sex + poly(age,2,raw=T) |
               t + city | (educ ~ policy), data)
coeftest(fit.did)

##
## t test of coefficients:
##
##                           Estimate  Std. Error  t value Pr(>|t|)
## sex                    -3.0334e-01  2.6932e-02 -11.2629  < 2e-16 ***
## poly(age, 2, raw = T)1  9.4895e-02  3.9460e-03  24.0483  < 2e-16 ***
## poly(age, 2, raw = T)2 -8.7531e-04  5.0545e-05 -17.3174  < 2e-16 ***
## `educ(fit)`             5.0000e-01  2.0394e-01   2.4518  0.01422 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Comparison

	Basic	OLS	Matching	IV	FE	FEIV	RDD	RDDX	DID
(Intercept)	10.221	8.138	10.209	8.264			10.462	−0.079	
	(0.008)	(0.224)	(0.013)	(0.229)			(0.016)	(0.015)	
educ	0.822	0.816	0.834	0.532	0.699	0.480	0.418	0.469	0.500
	(0.015)	(0.014)	(0.018)	(0.028)	(0.014)	(0.027)	(0.033)	(0.030)	(0.204)
sex		−0.277		−0.301	−0.286	−0.305			−0.303
		(0.013)		(0.014)	(0.013)	(0.013)			(0.027)
age		0.094		0.094	0.095	0.095			0.095
		(0.012)		(0.013)	(0.012)	(0.012)			(0.004)
age2		−0.001		−0.001	−0.001	−0.001			−0.001
		(0.000)		(0.000)	(0.000)	(0.000)			(0.000)

Going Beyond Linearity

- So far we have been dealing with a binary treatment variable and have been relying exclusively on linear statistical modeling for each of our causal inference designs.

- When the treatment variable has many levels or is continuous, its causal relation with the outcome variable is often nonlinear and the treatment effect is non-constant:

$$\text{ATE}\,(x) = \frac{d}{dx}\mathbb{E}\,[y|\text{do}(x)]$$

Going Beyond Linearity

- To the extent that a causal effect is heterogeneous, the ATE of a large population conveys limited information.
- To have a better understanding, we need to know how treatment effects vary among the population – how a given treatment will have different effects on different individuals.

Going Beyond Linearity

- Given a set of known effect modifiers s, we can estimate the **heterogeneous treatment effect** (**HTE**):

$$\text{ATE}\,(x, s) = \frac{\partial}{\partial x}\mathbb{E}\,[y|\text{do}(x), s]$$

- In a large socio-economic sample, the degree of heterogeneity can be substantial: individuals, firms, and markets vary in numerous ways. This requires the estimation of high-dimensional HTE.

Going Beyond Linearity

- Estimation of non-constant and (high-dimensional) heterogeneous causal effects requires sophisticated statistical models.
- Combining sound causal inference strategy (based on careful causal reasoning) with state-of-the-art statistical modeling allows us to most accurately estimate and predict causal effects of interest.

Structural Estimation

- **Causal models**, or **scientific models**, are mathematical models of causal mechanisms.

- In the econometrics literature, causal models based on economic theory are referred to as **structural models**. These models use economic theory to specify the functional forms of causal relationships.

Structural Estimation

- The estimation of structural models is called **structural estimation** – rather than learning a specific causal effect, structural estimation aims to estimate all the parameters of a causal model.

- In contrast, the use of statistical models for learning causal effects (based on given identification strategies) has been called **reduced-form analysis** in the econometrics literature[49].

[49] Historically, given a structural model $g(x, y) = 0$ that specifies the relationship governing exogenous variable x and endogenous variable y, if y is solved as a function of x, i.e. $y = f(x)$, then f is referred to as the **reduced form** of g.

Structural Estimation

Scientific vs. Statistical Model

If you want to predict where Mars will be in the night sky[a], you may do very well with a model in which Mars revolves around the Earth. You can estimate, from data, how fast Mars goes around the Earth and where it should be tonight. But the estimated model does not describe the actual causal mechanisms.

[a] Example taken from Shalizi (2016).

Structural Estimation

Returns to College Education

When we estimate the returns to college education in a reduced-form analysis, what are we estimating?

- Wage is an equilibrium outcome. How much wage a person would earn if she obtains a college degree depends on labor demand and labor supply. Labor supply, in turn, depends on how many other people have received college education[a].
- The effect of college education on wage is clearly different if only one person receives college education and if all individuals do.

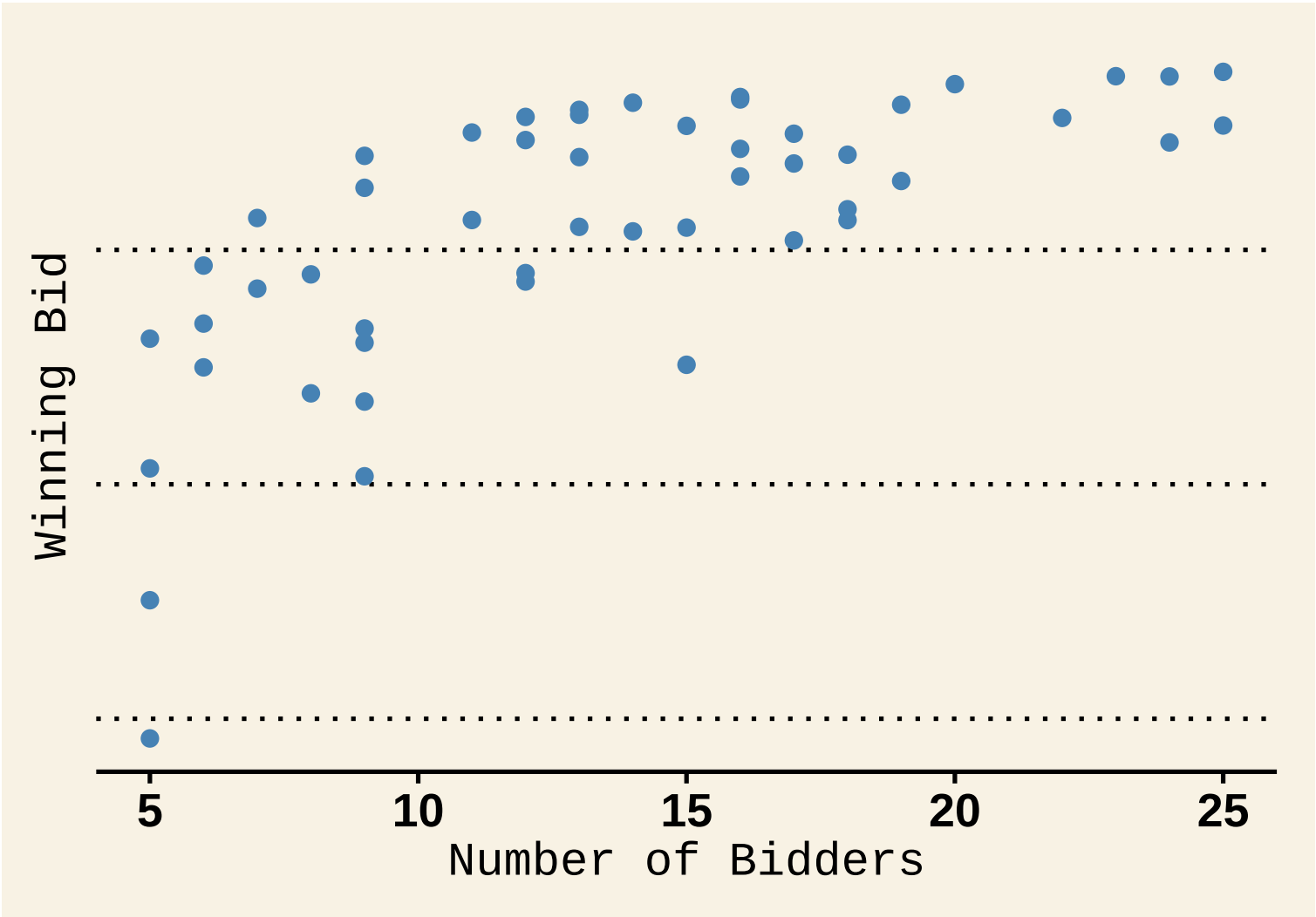
[a] Disregarding the heterogeneity in college education (good college, bad college, history major, economics major) and treating it as homogeneous here.

Structural Estimation

Returns to College Education

- The effect estimated in a reduced-form analysis is the effect of an individual receiving college education on her wage conditional on current labor demand and labor supply.
- Structural analysis would estimate labor demand and labor supply curves directly from data and compute the resulting equilibrium returns to education.
- Based on the estimated structural model, we can predict how returns to college education change when either demand or supply shifts due to exogenous forces (counterfactual prediction).

Auction



First-price Sealed-bid Auctions for a Given Good

Auction

Model

- N risk-neutral bidders
- Independent private value $v_i \sim^{i.i.d.} F(.)$
- Each bidder knows her own v_i and the distribution F, but not the v_i of others
- Observed bids are the Bayesian Nash equilibrium outcome of the game

Auction

Model

Equilibrium bidding strategy:

$$b_i = v_i - \frac{1}{F(v_i)^{N-1}} \int_0^{v_i} F(x)^{N-1} \, dx$$
$$= v_i - \frac{1}{N-1} \frac{G_N(b_i)}{g_N(b_i)}$$

, where $G_N(.)$ and $g_N(.)$ are the c.d.f. and p.d.f. of the bid distribution.

Auction

Structural Estimation

1. For each auction[a], nonparametrically estimate $G_N(.)$ and $g_N(.)$ from observed bids $\{b_1, \ldots, b_N\}$.
2. For each bidder, calculate

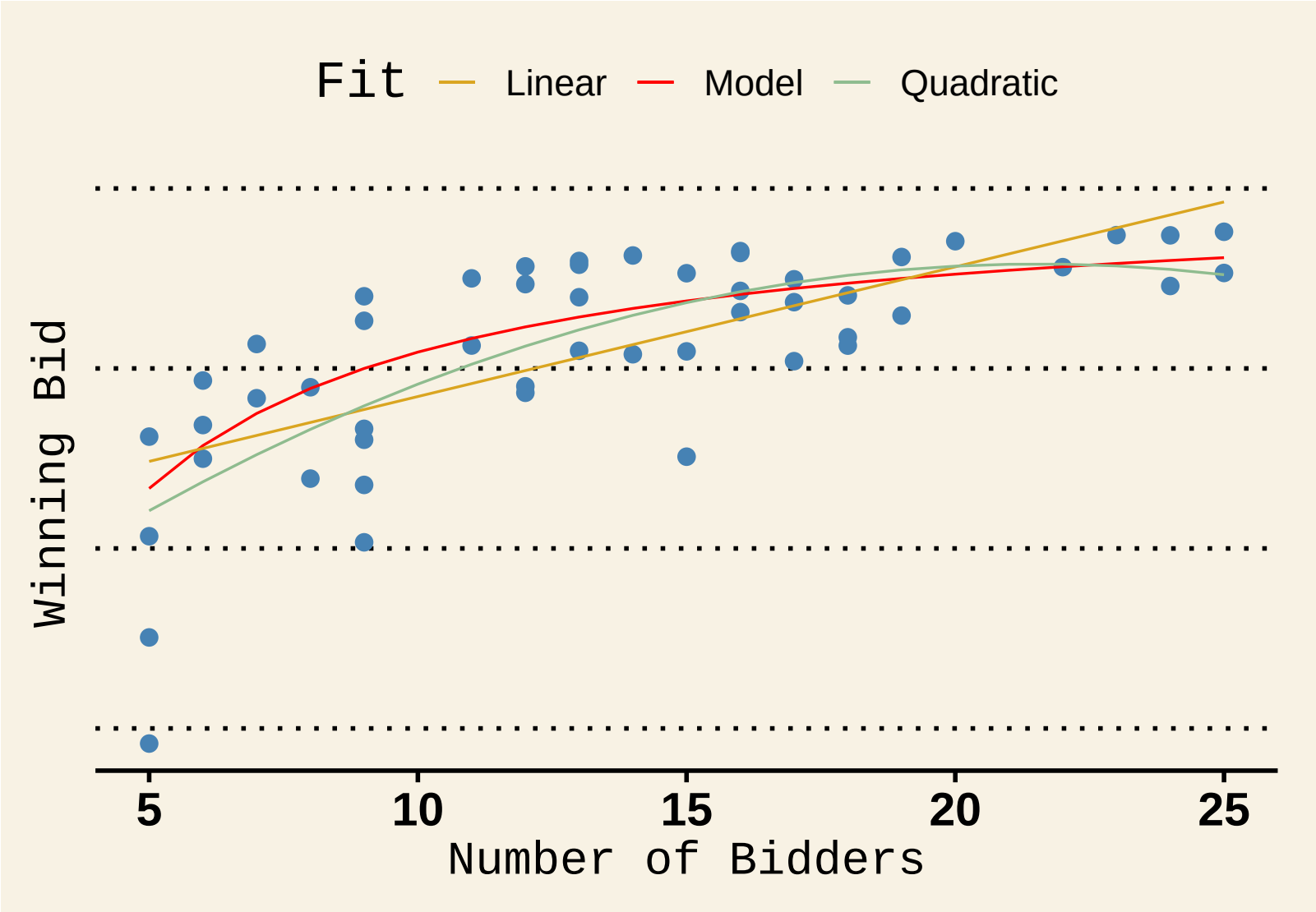
$$\widehat{v}_i = b_i + \frac{1}{N-1} \frac{\widehat{G}_N(b_i)}{\widehat{g}_N(b_i)} \tag{8}$$

3. Use $\widehat{v}_i$ to nonparametrically estimate $F(.)$
4. $\widehat{F}(.)$ can be used to predict the winning bid in an $N-$bidder auction:

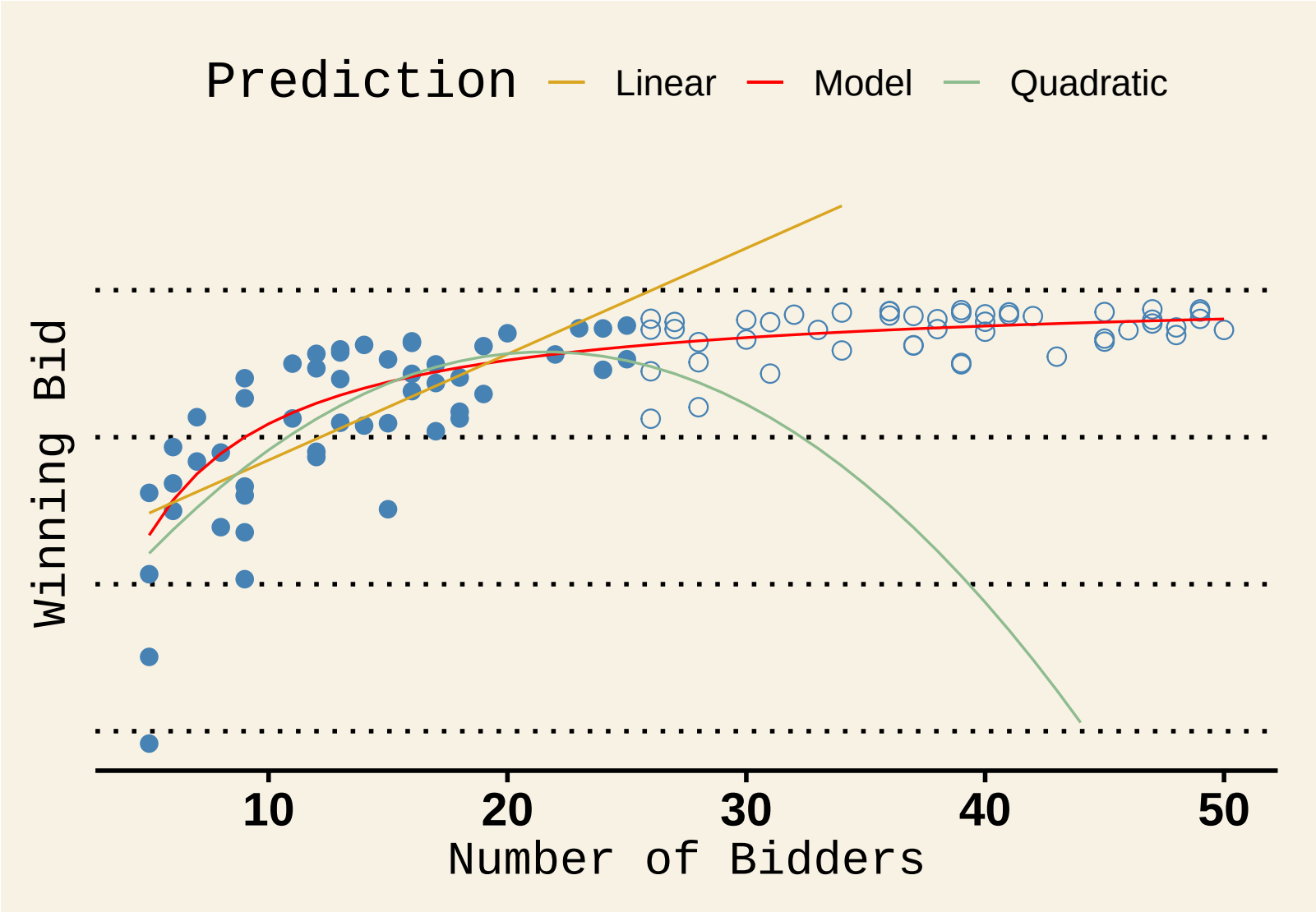
$$E[\max\{b_i\}] = E\left[\max\left\{v_i - \frac{1}{\widehat{F}(v_i)^{N-1}} \int_0^{v_i} \widehat{F}(x)^{N-1} dx\right\}\right]$$

[a] See Guerre et al. (2000).

Auction



Auction



Auction

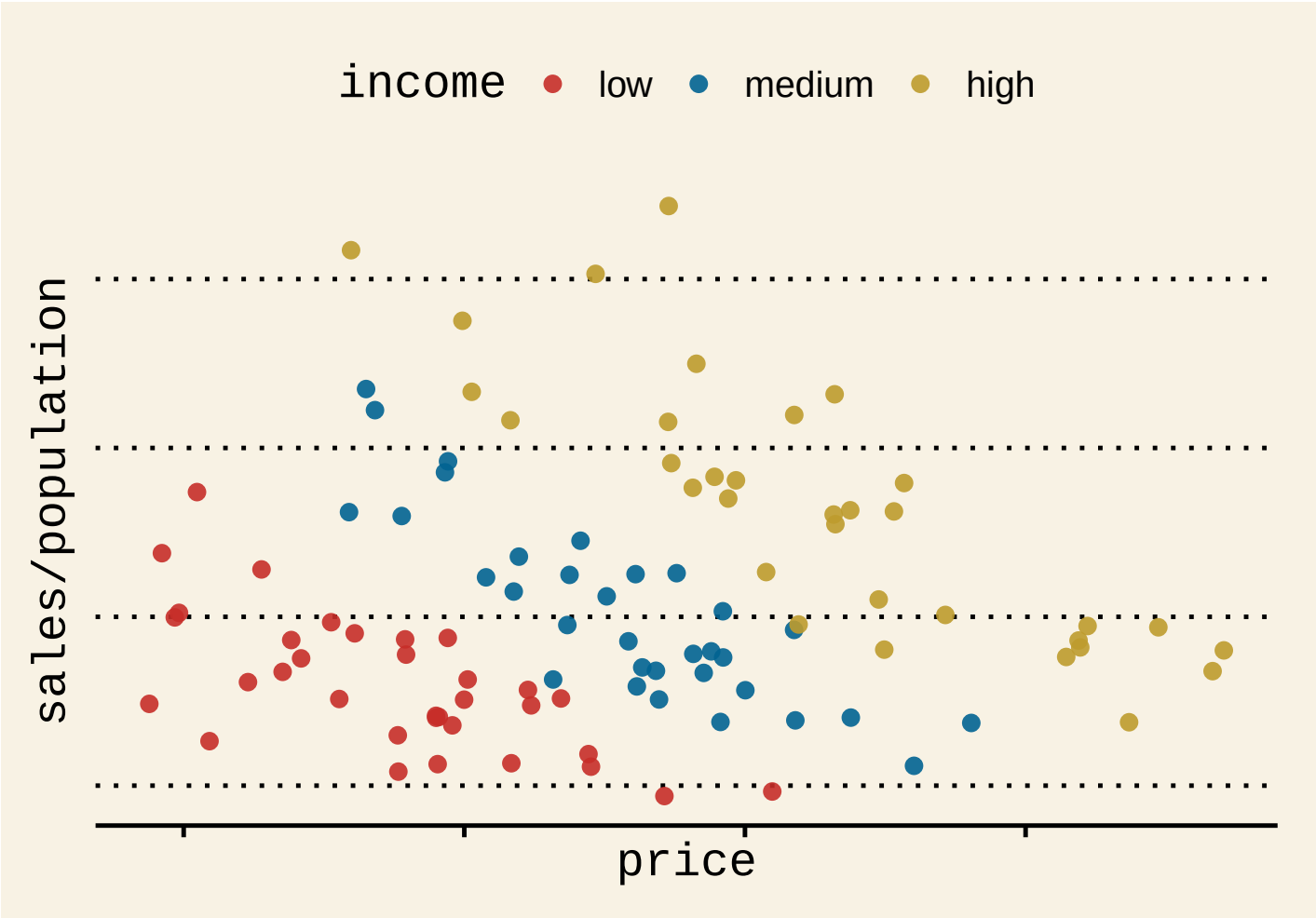
- We are interested in predicting the winning bid ($b_{\max}$) based on the number of bidders N. Let $f(N) \doteq \mathbb{E}[b_{\max} | N]$[50].
- The problem is to learn $f(N)$ from data. Here, theory helps specify the functional form of $f(N)$.
- Theory also helps us to learn the values of the bidders (equation (8)) by specifying the functional form of the mapping from $\{v_i\}$ to $\{b_i\}$.

[50] If the number of bidders is exogenous (such as in a randomized experiment), then $f(N)$ represents the causal effect of N on $b_{\max}$.

Monopoly

A monopoly firm's pricing and sales in different geographical markets

Data: price, sales, average income, population for each market



Monopoly

Model: Demand

In each market m with population N_m and mean income I_m, consumers choose between the monopoly product and an outside good. Individual utilities are given by:

$$\begin{aligned} U_{i0}^m &= \epsilon_{i0}^m \\ U_{i1}^m &= \beta_0 + \beta_1 I_m - \beta_2 p_m + \epsilon_{i1}^m \end{aligned} \tag{9}$$

, where (U_{i0}^m, U_{i1}^m) are respectively the indirect utilities of the outside good and the monopoly product, and $\epsilon_{ij}^m \sim \text{Gumbel}(0, 1)$.

(9) $\Rightarrow q_m \sim \text{Binomial}(N_m, \pi_m)$, where

$$\pi_m = \frac{\exp(\beta_0 + \beta_1 I_m - \beta_2 p_m)}{1 + \exp(\beta_0 + \beta_1 I_m - \beta_2 p_m)}$$

Monopoly

Model: Supply

For each market m, given demand $q_m(p)$, the monopoly firm chooses p to maximize:

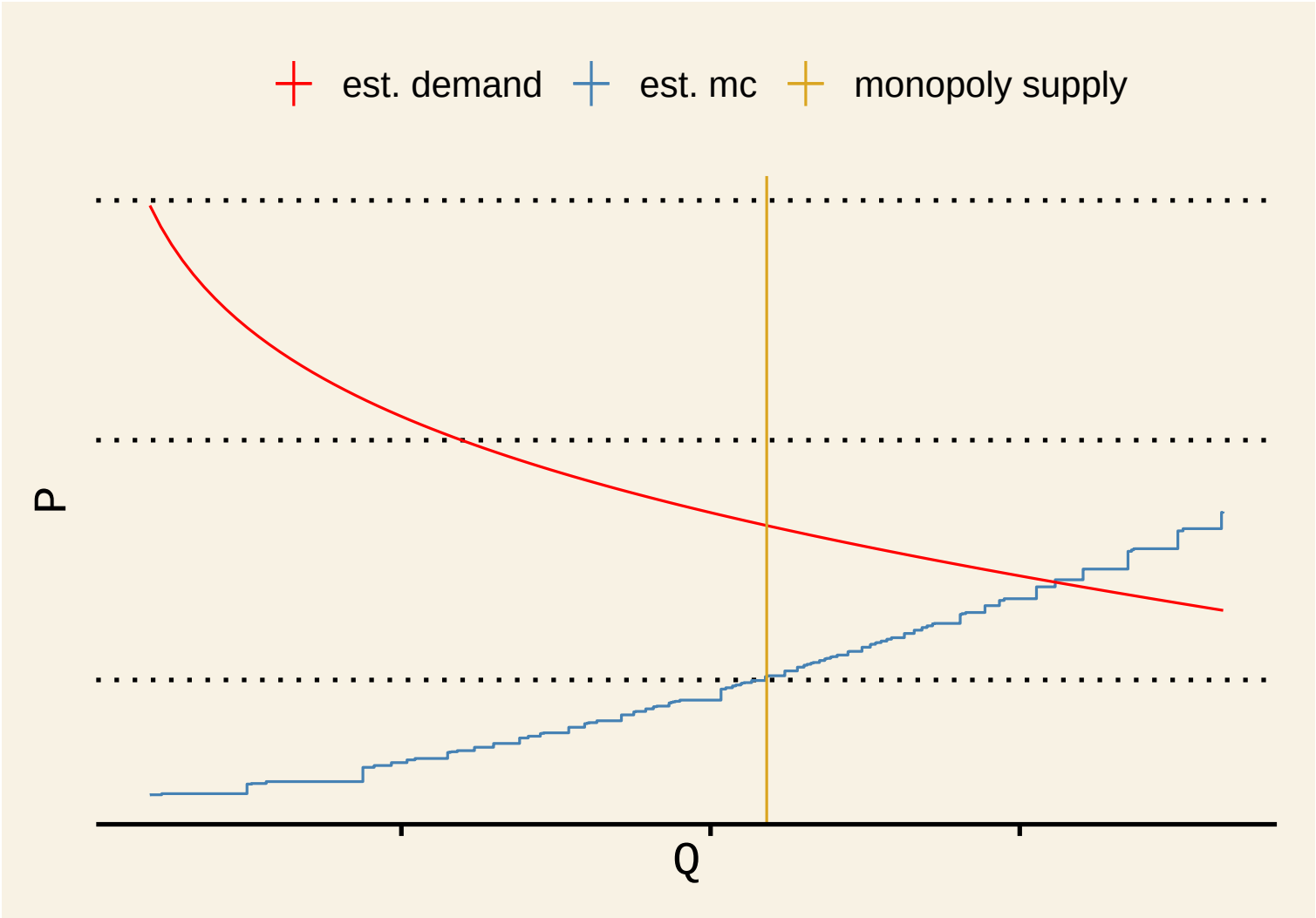
$$\max_p \{p \times q_m(p) - c(q_m(p))\} \tag{10}$$

, where $c(q)$ is the firm's cost function.

(10) $\Rightarrow$

$$c'(q_m) = p_m + \left[q'_m(p_m)\right]^{-1} q_m \tag{11}$$

Monopoly



Estimated marginal cost and demand curves
for a market with median income and population

Monopoly

- Here, theory helps us to learn the marginal cost function of the monopoly firm as well as the consumer utility function.
- Using the estimation results, we can conduct welfare analysis and make normative statements.
  - For example, calculating the total deadweight loss due to monopoly.

Counterfactual Simulation

- One of the benefits of learning a structural model is that it allows us to predict the effect of a completely new treatment – a treatment that has never been observed before.

- Once we have estimated a structural model with variables $\{x_1, \ldots, x_n\}$, we can use it to generate data from the distribution $p\left(x_1, \ldots, x_n \middle| \operatorname{do}\left(x_j = a\right)\right)$ for any hypothetical treatment $\operatorname{do}\left(x_j = a\right)$. This is called **counterfactual simulation**.

Counterfactual Simulation

What if Caesar never crossed the Rubicon?

Counterfactual Simulation

What happens if the government imposes a 20% sales tax on the monopoly firm?

After tax:

Δ `Consumer Surplus:`	-27.83%
Δ `Total Surplus:`	-27.95%

Tax incidence:

`Consumer:`	26.65%

Appendix

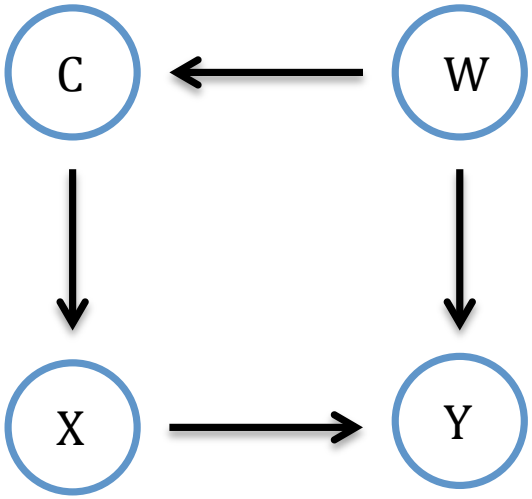


Figure 1:

Acknowledgement I

Part of this lecture is based on the following sources:

- Abu-Mostafa, Y. S., M. Magdon-Ismail, and H. Lin. 2012. *Learning from Data*. AMLBook.
- Angrist, J. D. and J. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Blei, D. M. *Interacting with data*. Lecture at Princeton University, retrieved on 2017.01.01. [link]
- Cunningham, C. (2021). *Causal Inference: The Mixtape*. Yale University Press.
- Doré, G. *The Dore Illustrations for Dante's Divine Comedy*. Dover Publications, 1st edition (1976).
- Hernán, M. A. and J. M. Robins. 2020. *Causal Inference*. CRC Press.

Acknowledgement II

- James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning: with Applications in R*. Springer.
- Silva, R. *Causal Inference in Machine Learning*. Talk at Imperial College London, retrieved on 2017.01.01. [link]
- Varian, H. R., *Machine Learning and Econometrics.* Talk at Google, retrieved on 2017.01.01. [link]

Reference

- Bajari, P., D. Nekipelov, S. P. Ryan, and M. Yang. 2015. "Machine Learning Methods for Demand Estimation," *American Economic Review*, 105(5).
- Guerre, E., I. Perrigne, and Q. Vuong. 2000. "Optimal Nonparametric Estimation of First-Price Auctions," *Econometrica*, 68(3).
- Shalizi, C. R. 2016. *Advanced Data Analysis from an Elementary Point of View*. Manuscript.