Introduction to Regression, or, "How I learned to Stop Worrying and Love Linear Regression"

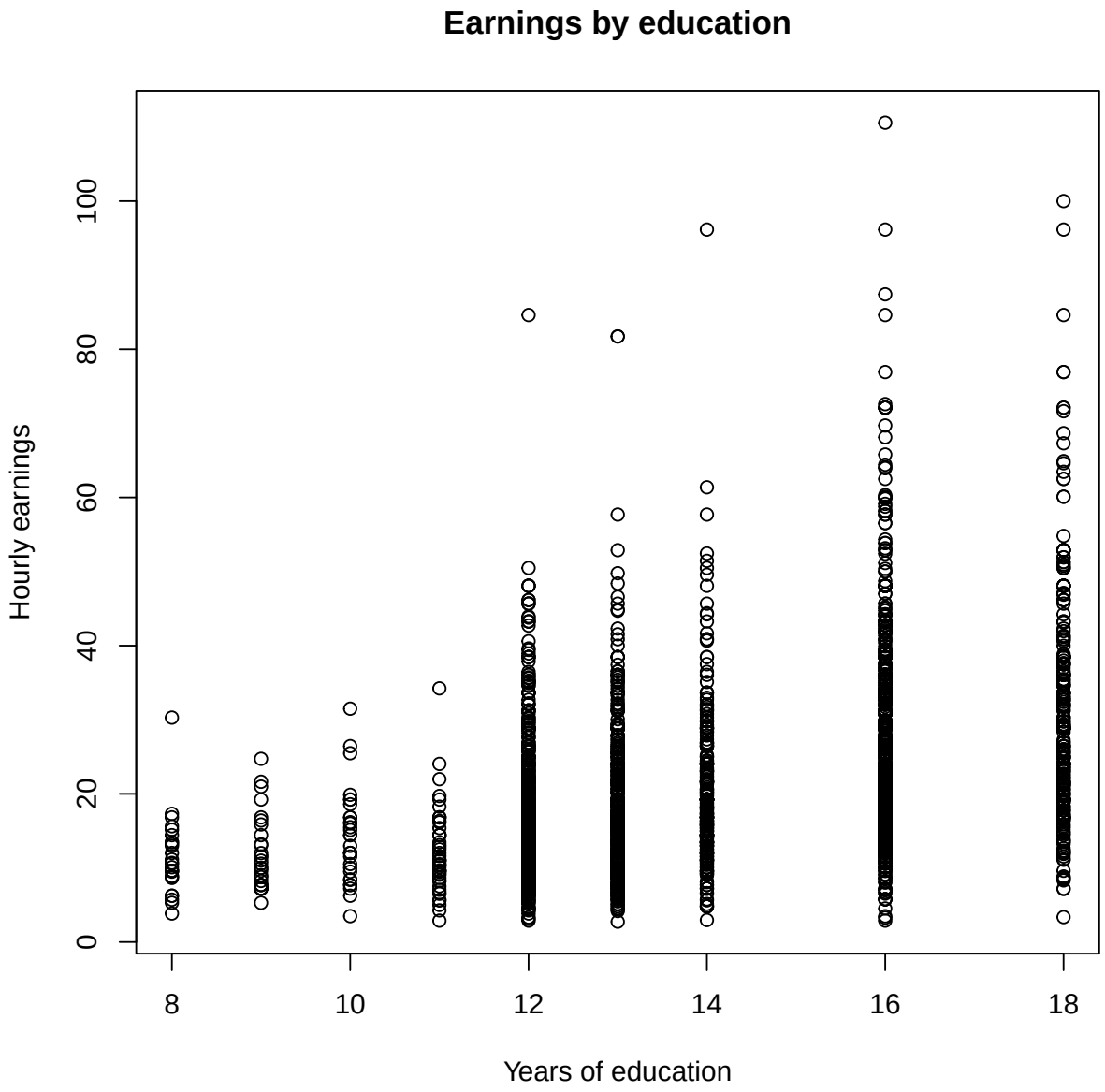
Econ 2560, Fall 2023

Prof. Josh Abel

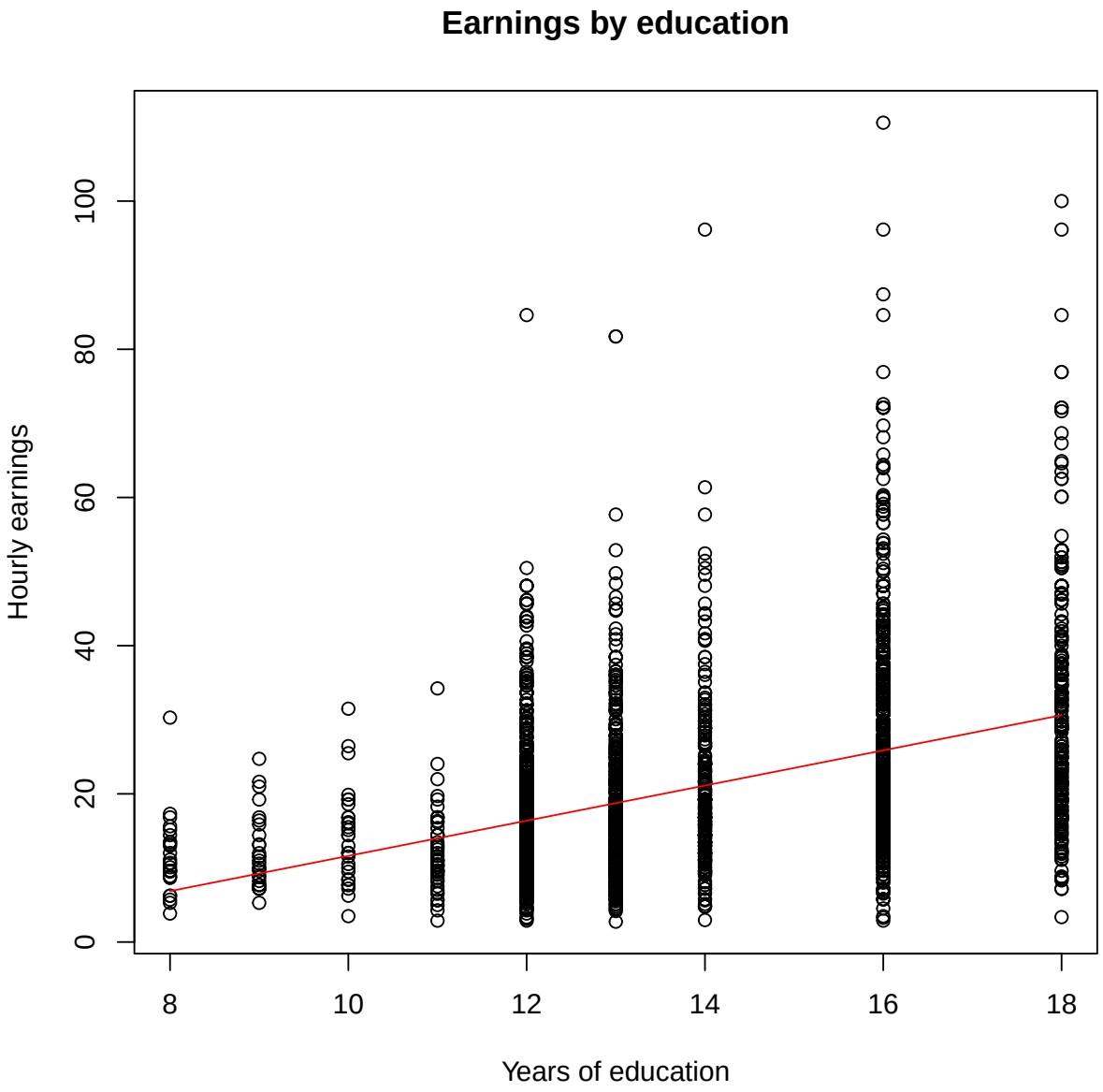
Introduction

- Regression: workhorse technique in empirical economic analysis
- Estimates mean of Y conditional on X ($E[Y|X]$)
- For now, focus on "univariate" model, where X is just one variable
- Often thought of as drawing a line "through the data"

A line through data



A line through data



Motivation

- We will learn how to draw the "best fit" line later
  - Ultimately, that's just some math
- Today we'll explore an alternative: **non-parametric regression**
  - This alternative is of increasing practical use, so helpful to know
  - Will also illustrate strengths and weaknesses of linear regression
- "Non-parametric regression" sounds fancy, but it is just about estimating means

Thought experiment

- Question: How much do Northeastern graduates earn at age 25, on average?
- Dataset:

ID	School	Age-25 Earnings
1	Northeastern	Y_1
...	...	...
5	Northeastern	Y_5
6	UMASS	Y_6
...	...	...
50	UMASS	Y_{50}

- How would you use this dataset to answer the question?

Thought experiment

- Question: How much do Northeastern graduates earn at age 25, on average?
- Dataset:

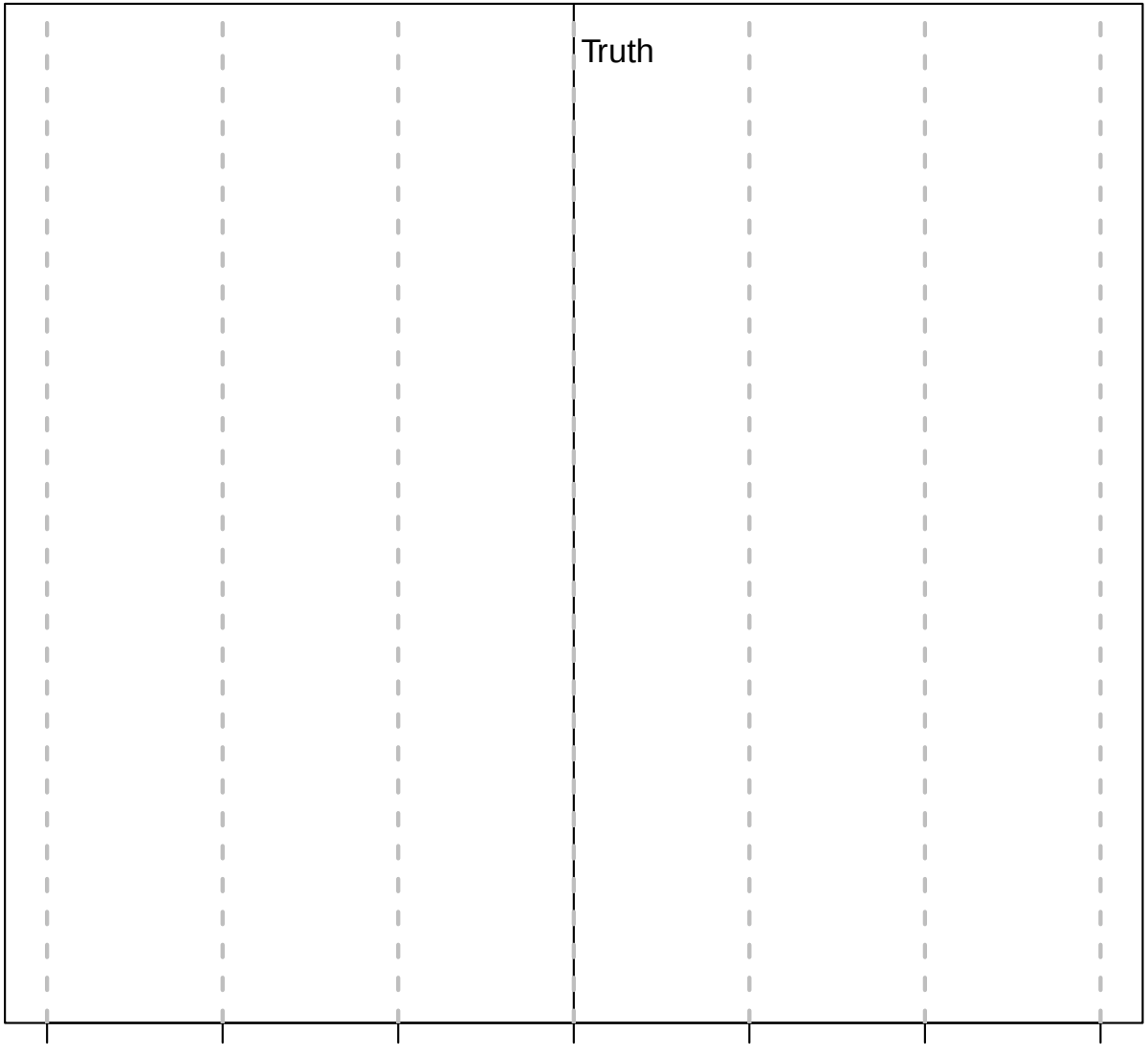
ID	School	Age-25 Earnings
1	Northeastern	Y_1
...	...	...
5	Northeastern	Y_5
6	UMASS	Y_6
...	...	...
50	UMASS	Y_{50}

- How would you use this dataset to answer the question?

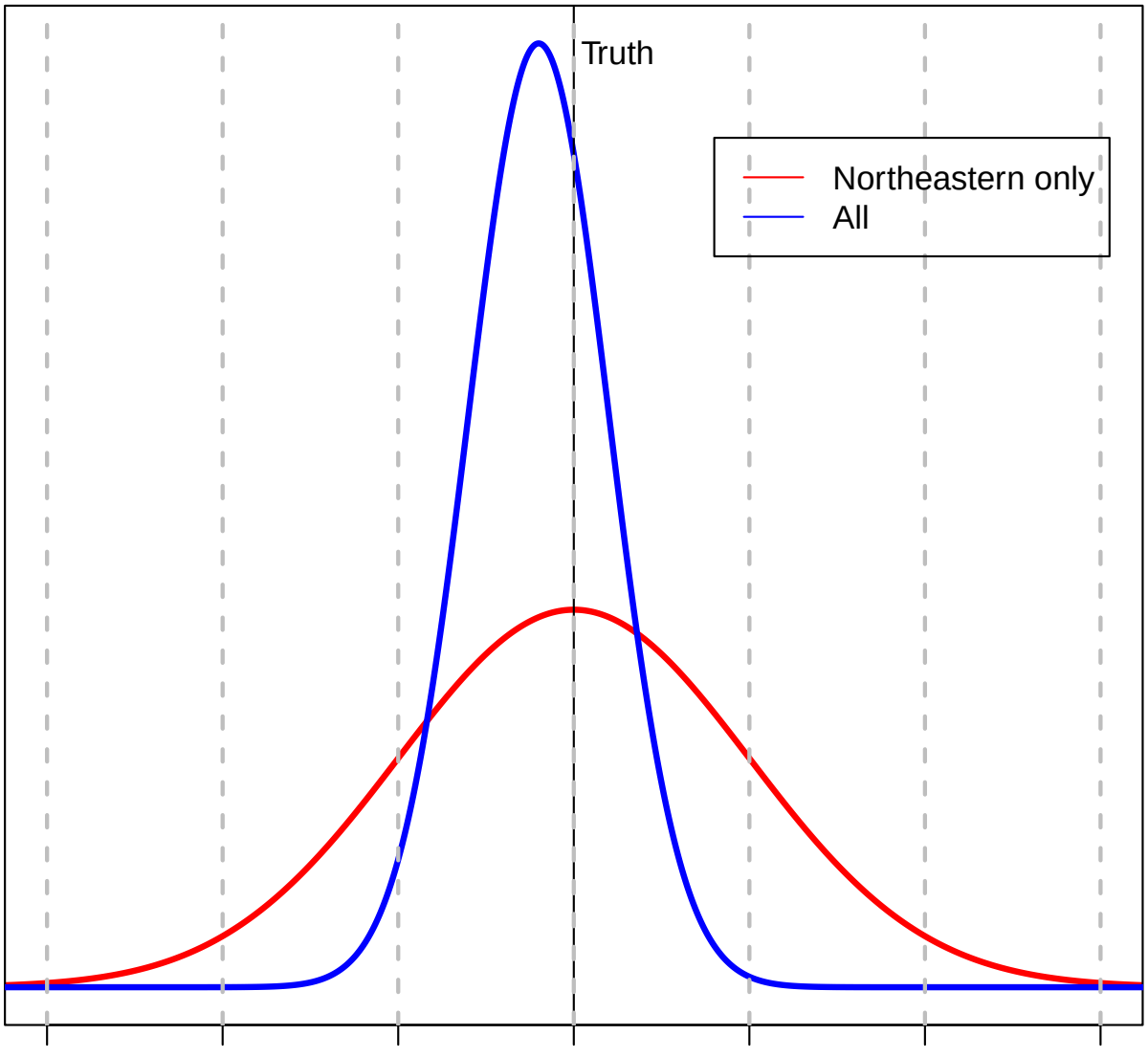
Option 1: $\frac{1}{5}\sum_{1}^{5} Y_i$

Option 2: $\frac{1}{50}\sum_{1}^{50} Y_i$

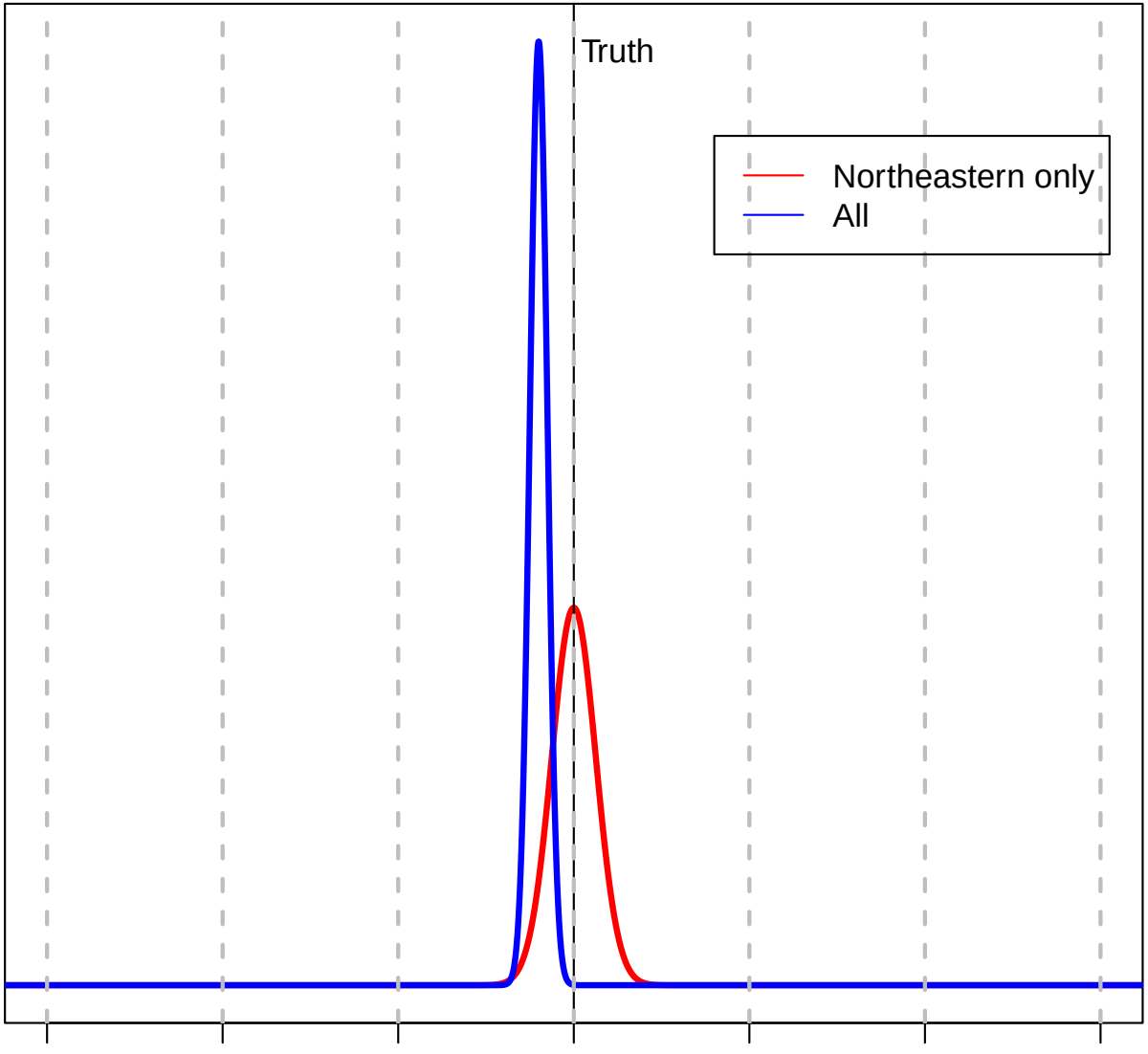
Graphical Intuition



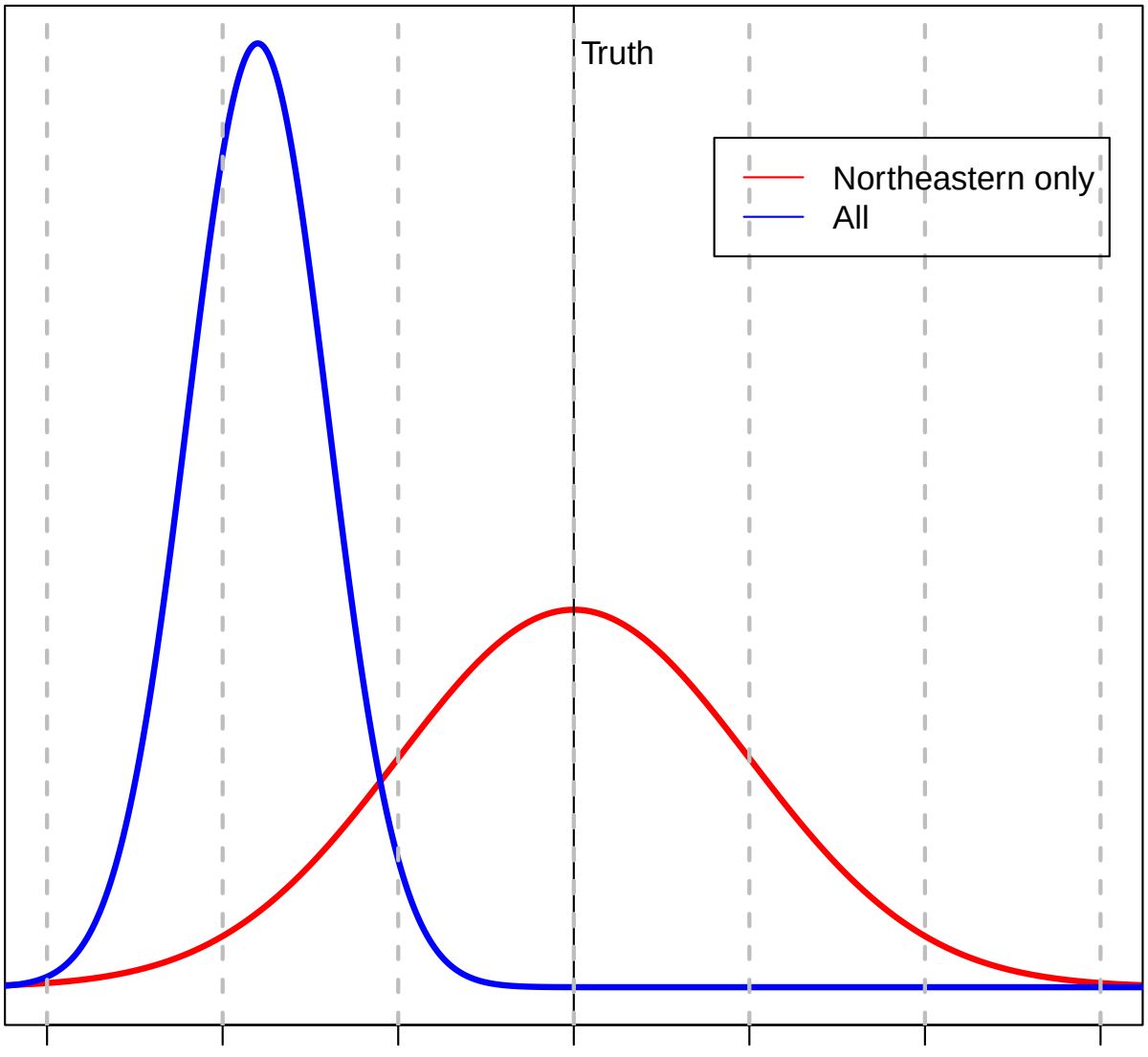
Biased approach may be more reliable...



...though not when the dataset is large...



...or if it's really biased



Bias-Variance Tradeoff

This intuition can be articulated mathematically

$$E\left[(\mu - \hat{\mu})^2\right] = \underbrace{(\mu - E[\hat{\mu}])^2}_{\text{Bias}^2} + \underbrace{E\left[(\hat{\mu} - E[\hat{\mu}])^2\right]}_{\text{Variance}}$$

μ: the "truth"/population mean
$\hat{\mu}$: estimator of μ/(function of data)

“So what?”

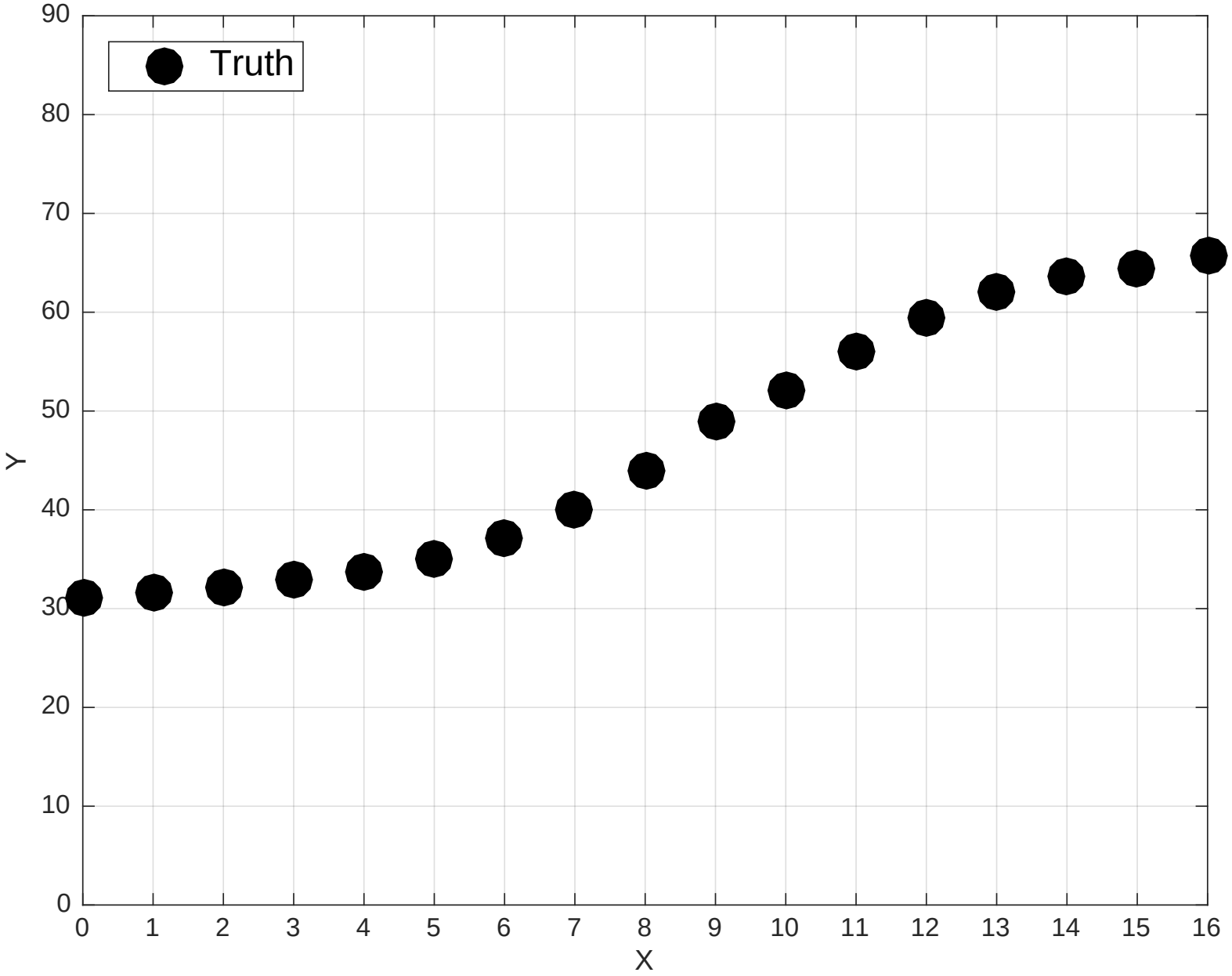
- Being pure/unbiased is not everything
- An unbiased estimator may be really erratic
- If introducing bias can reduce variance, it may be beneficial

Estimating a series of means

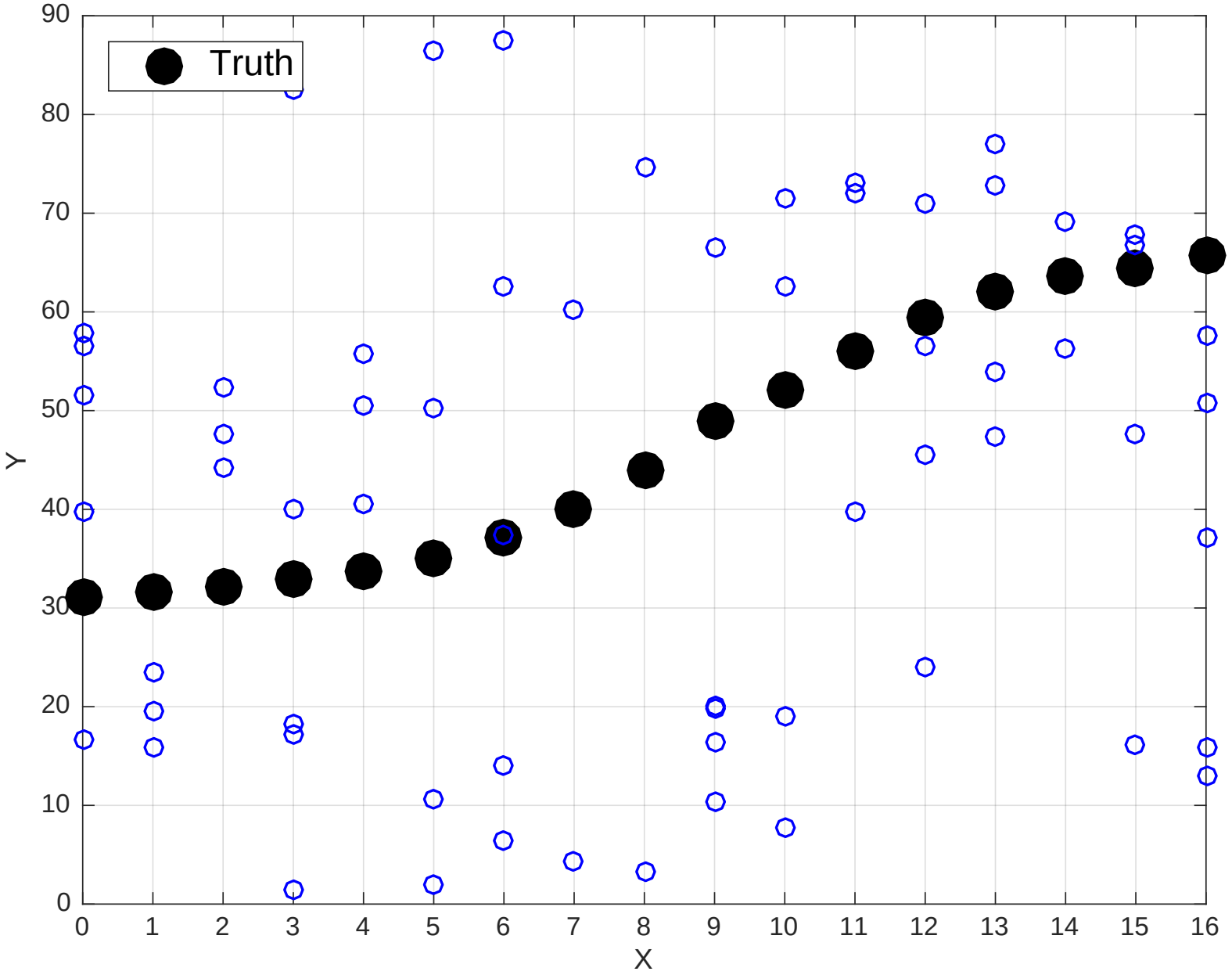
$$E[Y|X],\ X \in \{0, 1, 2, ..., K\}$$

For example, earnings (Y) as a function of years of education (X)

Estimating a series of means



Estimating a series of means



Estimating a series of means

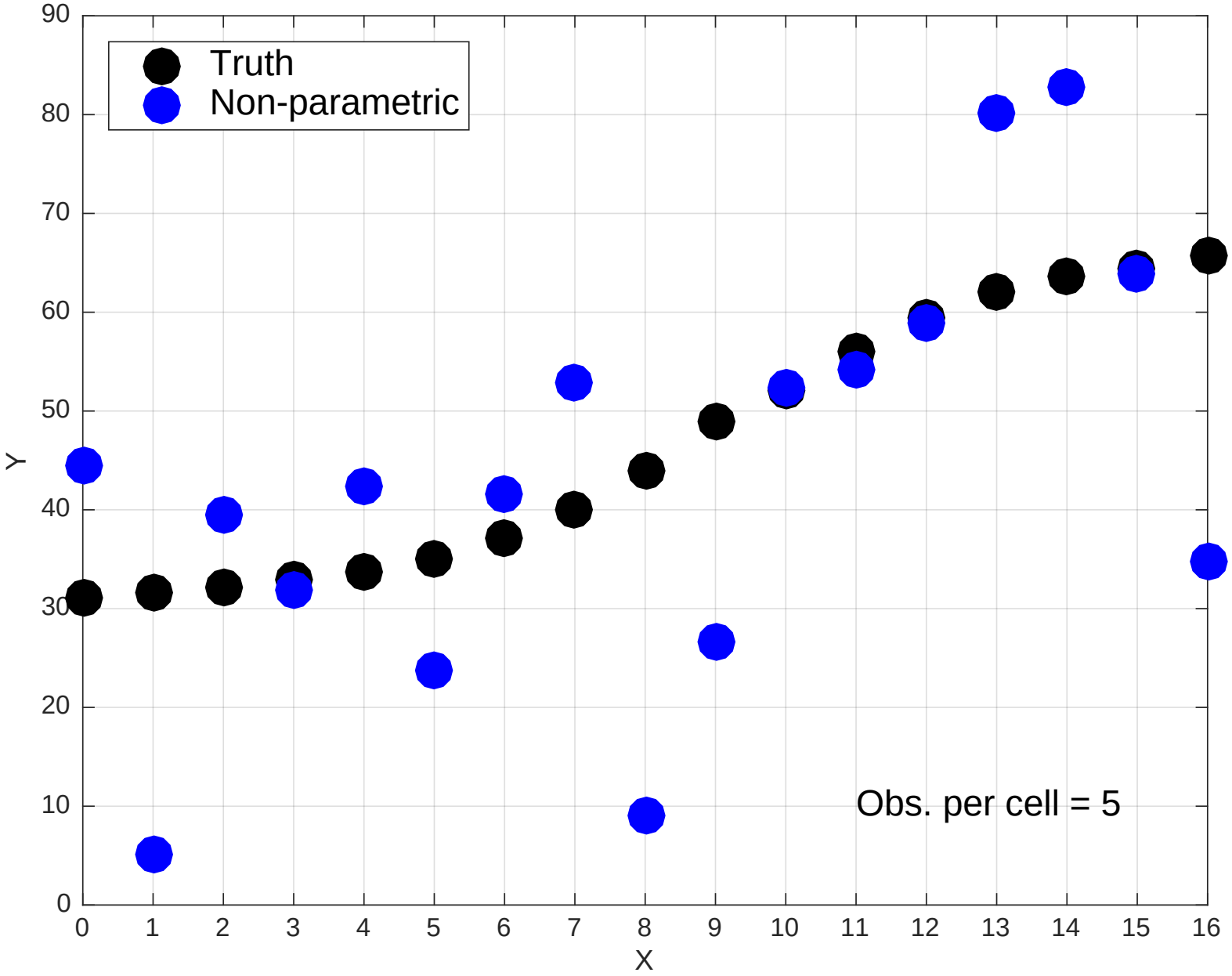
Linear specification/assumption:

$$E[Y|X] = \hat{\beta}_0 + \hat{\beta}_1 X$$

Assumption-free (**"non-parametric"**) specification:

$$E[Y|X] = \begin{cases} \bar{Y}_0 \text{ if } X = 0 \\ \bar{Y}_1 \text{ if } X = 1 \\ \\ \dots \\ \bar{Y}_K \text{ if } X = K \end{cases}$$

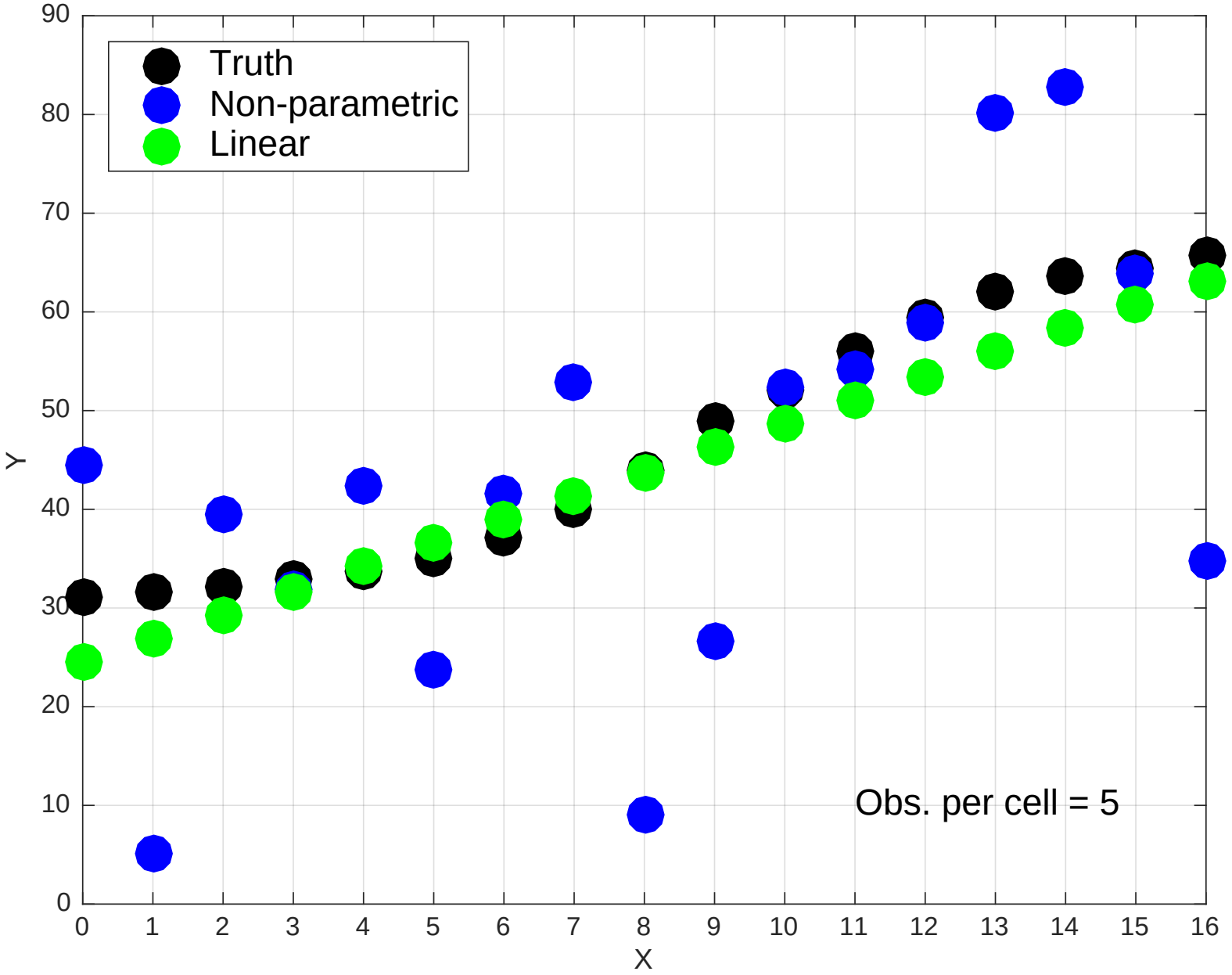
Non-parametric estimate on a "small" dataset



Non-parametric estimate on a "small" dataset

- Non-parametric approach is unbiased and consistent, point-wise
- But with our small dataset, the particular non-parametric estimate is erratic
  - Lots of ups and downs, some vary large
  - Not credible
- Too few observations for LLN to kick in

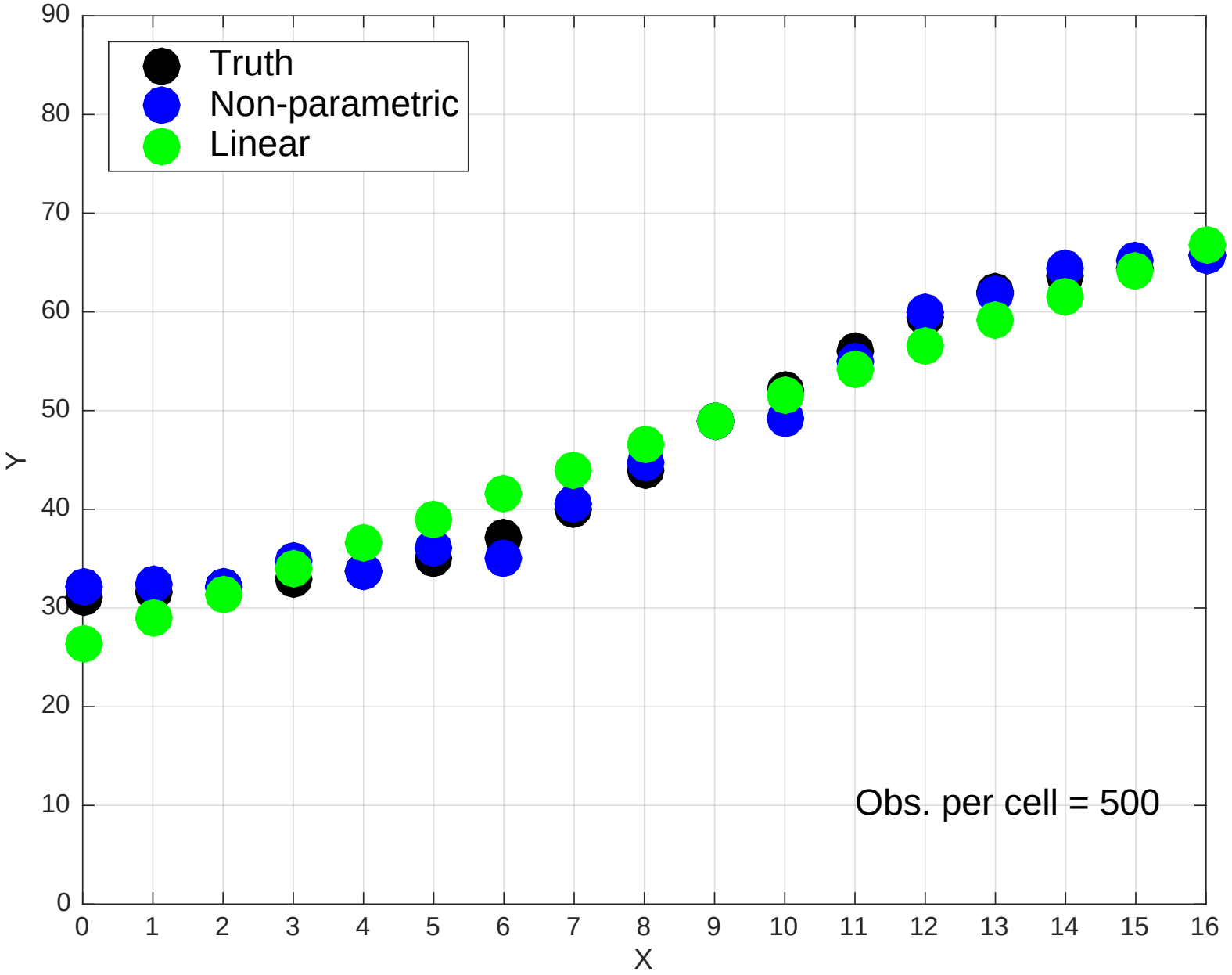
Linear estimate on a "small" dataset



Linear estimate on a "small" dataset

- Linear approach seems clearly preferable, though imperfect
- Linear approach is "biased"
  - Forces linear pattern, even though the truth is non-linear
- But in exchange for bias, we get stability
- Linear approach recognizes a pattern: *Educ* $\uparrow \rightarrow$ *Earnings* $\uparrow$
- It uses this to smooth over the bumpiness
- I.e. Because $\bar{Y}_{15} >> \bar{Y}_2$, will set $\hat{\mu}_5 > \hat{\mu}_4$, even though $\bar{Y}_5 < \bar{Y}_4$
  - Akin to using UMASS grads to estimate earnings of Northeastern grads

Estimates on a "large" dataset



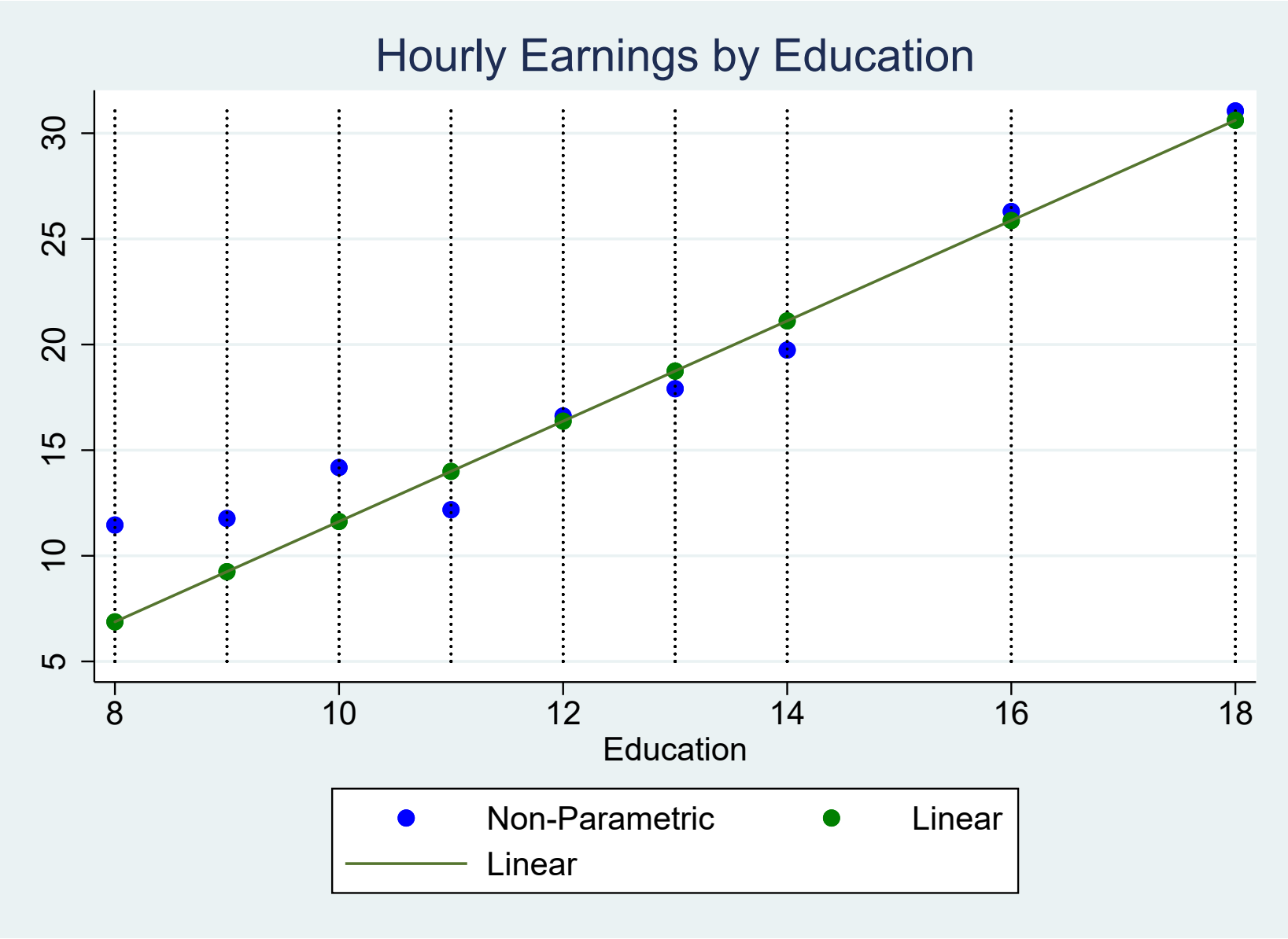
Estimates on a "large" dataset

- Non-parametric now seems better
- Now, even the "pure"/unbiased approach has enough data – not erratic
- The smoothing effect of a linearity assumption is now less helpful
  - Linear approach is still biased in that it can't capture non-linearity
  - No longer very helpful in giving stability, because non-parametric approach is already stable

Non-parametric regression: takeaways

- Non-parametric regression is a viable approach, particularly with large datasets
  - Imposes no assumptions – just relies on Law of Large Numbers!
- In smaller datasets, it may be erratic and unreliable
- Parametric (e.g. linear) regression may perform better in small datasets – better able to see the big picture and smooth things out
  - Can't get things exactly right
  - But you can be more confident it's not way off

With real data, which do you prefer?



With real data, which do you prefer?

- Non-parametric is not totally believable, because $\bar{Y}_{11} < \bar{Y}_{10}$
  - Could do a hypothesis test to see if they're statistically different!
- But linear cannot detect a discrete jump from 11 to 12 (high school graduation)
- This demonstrates precisely the tension between the two approaches
- Also demonstrates why econometrics is as much art as science

With real data, which do you prefer?

- Non-parametric is not totally believable, because $\bar{Y}_{11} < \bar{Y}_{10}$
  - Could do a hypothesis test to see if they're statistically different!
- But linear cannot detect a discrete jump from 11 to 12 (high school graduation)
- This demonstrates precisely the tension between the two approaches
- Also demonstrates why econometrics is as much art as science
- **There is no such thing as a "correct specification" – there are always tradeoffs!**

Moving forward with regression

- Moving forward, we will mostly focus on parametric (typically linear) approaches
- Why?
  - Non-parametric regression is simple – you already know how to estimate means!
    - It's a bit more subtle with continuous variables, but not too bad
  - Even large datasets can start to feel small when you include many variables in your model
    - "Curse of dimensionality": how many observations do you have of females in Connecticut with 10 years of education, etc.
  - Parametric approach gives a nice summary
    - E.g. the average effect of another year of education (2.4 \$/hr)
- But in real data work, don't forget about non-parametric regression!